IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

DECLARATION OF VISHWANATH R. IYER, Ph.D.
UNDER 37 C.F.R. § 1.132

I, VISHWANATH R. IYER, Ph.D., declare and state as follows:

1. I am an Assistant Professor in the Section of Molecular Genetics and Microbiology, Institute of Cellular and Molecular Biology, University of Texas at Austin, where my laboratory currently studies global transcriptional control in yeast, gene expression programs during human cell proliferation, and genome-wide transcription factor targets in yeast and human. Immediately prior to this position, I spent four years as a postdoctoral fellow in the laboratory of Patrick O. Brown at Stanford University studying the transcriptional programs of yeast and of human cells. My *curriculum vitae* is attached hereto as Exhibit A.

2. Beginning in Dr. Brown's laboratory, where I helped to develop the first whole genome arrays for yeast and early versions of highly representative cDNA arrays for human cells, and continuing to the present day, I have used microarray-based gene expression analysis as a principal approach in much of my research.

3. Representative publications describing this work include:

DeRisi J. *et al.*, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science* 278:680-686 (1997);[1]

Marton *et al.*, "Drug target validation and identification of secondary drug target effects using DNA microarrays," Nature Med. 4:1293-1301 (1998);[2]

Iyer *et al.*, "The transcriptional program in the response of human fibroblasts to serum," Science 283:83-87 (1999);[3] and

Ross *et al.*, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics* 24: 227-235 (2000).[4]

Two of the papers describe our use of microarray-based expression profiling to explore the metabolic reprogramming that occurs during major physiological changes, both in yeast (DeRisi *et al.*, during the shift from fermentation to respiration) and in human cells (Iyer *et al.*, human fibroblasts exposed to serum). One reference describes our use of expression profile analysis in drug target validation and identification of secondary drug effects (Marton *et al.*). And one describes our use of expression profiling as a molecular phenotyping tool to discriminate among human cancer cells (Ross *et al.*).

4. Whether used to elucidate basic physiological responses, to study primary and secondary drug effects, or to discriminate and classify human cancers, expression profiling

[1] Attached hereto as Exhibit B.

[2] Attached hereto as Exhibit C.

[3] Attached hereto as Exhibit D.

[4] Attached hereto as Exhibit E.

as we have practiced it relies for its power on comparison of ***patterns*** of expression.

5. For example, we have demonstrated that we can use the presence or absence of a characteristic drug "signature" pattern of altered gene expression in drug-treated cells to explore the mechanism of drug action, and to identify secondary effects that can signal potentially deleterious drug side effects. As another example, we have demonstrated that gene expression patterns can be used to classify human tumor cell lines. While it is of course advantageous to know the biological function of the encoded gene products in order to reach a better understanding of the cellular mechanisms underlying these results, these pattern-based analyses do not require knowledge of the biological function of the encoded proteins.

6. The resolution of the patterns used in such comparisons is determined by the number of genes detected: the greater the number of genes detected, the higher the resolution of the pattern. It goes without saying that higher resolution patterns are generally more useful in such comparisons than lower resolution patterns. With such higher resolutions comes a correspondingly higher degree of statistical confidence for distinguishing different patterns, as well as identifying similar ones.

7. Each gene included as a probe on a microarray provides a signal that is specific to the cognate transcript, at least to a first approximation.[5] Each new gene-specific

[5] In a more nuanced view, it is certainly possible for a probe to signal the presence of a variety of splice variants of a single gene,

(Continued...)

probe added to a microarray thus increases the number of genes detectable by the device, increasing the resolving power of the device. As I note above, higher resolution patterns are generally more useful in comparisons than lower resolution patterns. Accordingly, each new gene probe added to a microarray increases the usefulness of the device in gene expression profiling analyses. This proposition is so well-established as to be virtually an axiom in the art, and has been as long as I have been working in the field, and certainly since the time I embarked on the production of whole genome arrays in early 1996. Simply put, arrays with fewer gene-specific probes are inferior to arrays with more gene-specific probes.

8. For example, our ability to subdivide cancers into discriminable classes by expression profiling is limited by the resolution of the patterns produced. With more genes contributing to the expression patterns, we can potentially draw finer distinctions among the patterns, thus subdividing otherwise indistinguishable cancers into a greater number of classes; the greater the number of classes, the greater the likelihood that the cancers classified together will respond similarly to therapeutic intervention, permitting better individualization of therapy and, we hope, better treatment outcomes.

9. If a gene does not change expression in an experiment, or if a gene is not expressed and produces no

(...Continued)
without discriminating among them, and for a probe to signal the presence of a variety of allelic variants of a single gene, again without discriminating among them.

signal in an experiment, that is not to say that the probe lacks usefulness on the array; it only means that an insufficient number of conditions have been sampled to identify expression changes. In fact, an experiment showing that a gene is not expressed or that its expression level does not change can be equally informative. To provide maximum versatility as a research tool, the microarray should include -- and as a biologist I would want my microarray to include -- each newly identified gene as a probe.

10. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and may jeopardize the validity of any patent application in which this declaration is filed or any patent that issues thereon.

Vishwanath

VISHWANATH R. IYER, Ph.D.

October 20, 2003

Date

EXHIBIT A

Reference 8 of 9
with Response dated 4/23/04
In USSN: 10/049,742

SeqServer®
biology *in silico*

BLAST2 Search Results

Sequences | Help

Retrieval | BLAST2 | FASTA | ClustalW | GCG Assembly | Phrap | Translation | BLAST2 Manual

Confidential -- Property of Incyte Corporation   SeqServer Version 4.6 Jan 2002

Program: blastp
Sequence ID(s):

□ 1459372CD1 vs. genpept137

```
NCBI-BLASTP 2.0.10 [Aug-26-1999]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= 1459372CD1
         (269 letters)

Database: genpept137
           1,534,369 sequences; 474,463,515 total letters

Searching..................................................done

                                                                   Score     E
Sequences producing significant alignments:                        (bits)  Value

☑ g16550798   unnamed protein product [Homo sapiens]                   562   e-159
☑ g14194055   dopamine receptor interacting protein [Rattus norve      536   e-151
☑ g15777195   J-domain protein Jiv [Bos taurus]                        535   e-151
☑ g15777193   J-domain protein Jiv [Bos taurus]                        535   e-151
☑ g26337373   unnamed protein product [Mus musculus]                   534   e-150
☑ g12857284   unnamed protein product [Mus musculus]                   534   e-150
☑ g15843561   DnaJ1 protein [Bos taurus]                               533   e-150
☑ g15029846   RIKEN cDNA 5730551F12 gene [Mus musculus]                533   e-150
☑ g26349793   unnamed protein product [Mus musculus]                   531   e-150
☑ g26337271   unnamed protein product [Mus musculus]                   454   e-126

>g16550798 unnamed protein product [Homo sapiens]
          Length = 412

 Score =  562 bits (1433), Expect = e-159
 Identities = 269/269 (100%), Positives = 269/269 (100%)

Query: 1    MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 60
             MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI
```

```
Sbjct: 144 MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 203

Query: 61  VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 120
           VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP
Sbjct: 204 VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 263

Query: 121 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 180
           KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH
Sbjct: 264 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 323

Query: 181 RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAPGA 240
           RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAPGA
Sbjct: 324 RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAPGA 383

Query: 241 AAASKPNSTVPKGEAKPKRRKKVRRPFQR 269
           AAASKPNSTVPKGEAKPKRRKKVRRPFQR
Sbjct: 384 AAASKPNSTVPKGEAKPKRRKKVRRPFQR 412


>g14194055 dopamine receptor interacting protein [Rattus norvegicus]
           Length = 701

 Score =  536 bits (1365), Expect = e-151
 Identities = 252/269 (93%), Positives = 260/269 (95%)

Query: 1   MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 60
           MAGVPEDELNPFHVLGVEATASD+ELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI
Sbjct: 433 MAGVPEDELNPFHVLGVEATASDIELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 492

Query: 61  VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 120
           VSN E+RKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP
Sbjct: 493 VSNPERRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 552

Query: 121 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 180
           KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH
Sbjct: 553 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 612

Query: 181 RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAPGA 240
           RVPYHISFGSR+PGT GRQRATP++PPADLQDFLSRIFQVPPG M NGNFFAAP P PG
Sbjct: 613 RVPYHISFGSRVPGTSGRQRATPESPPADLQDFLSRIFQVPPGPMSNGNFFAAPHPGPGT 672

Query: 241 AAASKPNSTVPKGEAKPKRRKKVRRPFQR 269
            + S+PNS+VPKGEAKPKRRKKVRRPFQR
Sbjct: 673 TSTSRPNSSVPKGEAKPKRRKKVRRPFQR 701


>g15777195 J-domain protein Jiv [Bos taurus]
           Length = 699

 Score =  535 bits (1364), Expect = e-151
 Identities = 258/269 (95%), Positives = 260/269 (95%), Gaps = 4/269 (1%)

Query: 1   MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 60
           MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI
Sbjct: 435 MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 494

Query: 61  VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 120
           VSN E+RKEYEMKRMAENELSRSVNEFLSKLQ    EAMNTMMCSRCQGKHRRFEMDREP
Sbjct: 495 VSNPERRKEYEMKRMAENELSRSVNEFLSKLQ----EAMNTMMCSRCQGKHRRFEMDREP 550

Query: 121 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 180
           KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH
Sbjct: 551 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 610

Query: 181 RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAPGA 240
           RVPYHISFGSR+PGT GRQRATPDAPPADLQDFLSRIFQVPPGQM NGNFFAAPQP PGA
Sbjct: 611 RVPYHISFGSRMPGTSGRQRATPDAPPADLQDFLSRIFQVPPGQMSNGNFFAAPQPGPGA 670
```

```
Query: 241 AAASKPNSTVPKGEAKPKRRKKVRRPFQR 269
            AASKPNSTVPKGEAKPKRRKKVRRPFQR
Sbjct: 671 TAASKPNSTVPKGEAKPKRRKKVRRPFQR 699


>g15777193 J-domain protein Jiv [Bos taurus]
           Length = 699

 Score =  535 bits (1364), Expect = e-151
 Identities = 258/269 (95%), Positives = 260/269 (95%), Gaps = 4/269 (1%)

Query: 1   MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 60
           MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI
Sbjct: 435 MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 494

Query: 61  VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 120
           VSN E+RKEYEMKRMAENELSRSVNEFLSKLQ    EAMNTMMCSRCQGKHRRFEMDREP
Sbjct: 495 VSNPERRKEYEMKRMAENELSRSVNEFLSKLQ----EAMNTMMCSRCQGKHRRFEMDREP 550

Query: 121 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 180
           KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH
Sbjct: 551 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 610

Query: 181 RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAPGA 240
           RVPYHISFGSR+PGT GRQRATPDAPPADLQDFLSRIFQVPPGQM NGNFFAAPQP PGA
Sbjct: 611 RVPYHISFGSRMPGTSGRQRATPDAPPADLQDFLSRIFQVPPGQMSNGNFFAAPQPGPGA 670

Query: 241 AAASKPNSTVPKGEAKPKRRKKVRRPFQR 269
            AASKPNSTVPKGEAKPKRRKKVRRPFQR
Sbjct: 671 TAASKPNSTVPKGEAKPKRRKKVRRPFQR 699


>g26337373 unnamed protein product [Mus musculus]
           Length = 703

 Score =  534 bits (1361), Expect = e-150
 Identities = 251/269 (93%), Positives = 259/269 (95%)

Query: 1   MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 60
           MAGVPEDELNPFHVLGVEATASD ELKKAYRQLAVMVHPDKNHHPRAEEAFK+LRAAWDI
Sbjct: 435 MAGVPEDELNPFHVLGVEATASDTELKKAYRQLAVMVHPDKNHHPRAEEAFKILRAAWDI 494

Query: 61  VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 120
           VSN E+RKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP
Sbjct: 495 VSNPERRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 554

Query: 121 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 180
           KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH
Sbjct: 555 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 614

Query: 181 RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAPGA 240
           RVPYHISFGSR+PGT GRQRATP++PPADLQDFLSRIFQVPPG M NGNFFAAP P PG
Sbjct: 615 RVPYHISFGSRVPGTSGRQRATPESPPADLQDFLSRIFQVPPGPMSNGNFFAAPHPGPGT 674

Query: 241 AAASKPNSTVPKGEAKPKRRKKVRRPFQR 269
            + S+PNS+VPKGEAKPKRRKKVRRPFQR
Sbjct: 675 TSTSRPNSSVPKGEAKPKRRKKVRRPFQR 703


>g12857284 unnamed protein product [Mus musculus]
           Length = 703

 Score =  534 bits (1361), Expect = e-150
 Identities = 251/269 (93%), Positives = 259/269 (95%)

Query: 1   MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 60
```

```
           MAGVPEDELNPFHVLGVEATASD ELKKAYRQLAVMVHPDKNHHPRAEEAFK+LRAAWDI
Sbjct: 435 MAGVPEDELNPFHVLGVEATASDTELKKAYRQLAVMVHPDKNHHPRAEEAFKILRAAWDI 494

Query: 61  VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 120
           VSN E+RKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP
Sbjct: 495 VSNPERRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 554

Query: 121 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 180
           KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH
Sbjct: 555 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 614

Query: 181 RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAPGA 240
           RVPYHISFGSR+PGT GRQRATP++PPADLQDFLSRIFQVPPG M NGNFFAAP P PG
Sbjct: 615 RVPYHISFGSRVPGTSGRQRATPESPPADLQDFLSRIFQVPPGPMSNGNFFAAPHPGPGT 674

Query: 241 AAASKPNSTVPKGEAKPKRRKKVRRPFQR 269
            + S+PNS+VPKGEAKPKRRKKVRRPFQR
Sbjct: 675 TSTSRPNSSVPKGEAKPKRRKKVRRPFQR 703


>g15843561 DnaJ1 protein [Bos taurus]
           Length = 659

 Score =  533 bits (1358), Expect = e-150
 Identities = 257/269 (95%), Positives = 259/269 (95%), Gaps = 4/269 (1%)

Query: 1   MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 60
           MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI
Sbjct: 395 MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 454

Query: 61  VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 120
           VSN E+RKEYEMKRMAENELSRSVNEFLSKLQ    EAMNTMMCSRCQGKHR FEMDREP
Sbjct: 455 VSNPERRKEYEMKRMAENELSRSVNEFLSKLQ----EAMNTMMCSRCQGKHRSFEMDREP 510

Query: 121 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 180
           KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH
Sbjct: 511 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 570

Query: 181 RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAPGA 240
           RVPYHISFGSR+PGT GRQRATPDAPPADLQDFLSRIFQVPPGQM NGNFFAAPQP PGA
Sbjct: 571 RVPYHISFGSRMPGTSGRQRATPDAPPADLQDFLSRIFQVPPGQMSNGNFFAAPQPGPGA 630

Query: 241 AAASKPNSTVPKGEAKPKRRKKVRRPFQR 269
            AASKPNSTVPKGEAKPKRRKKVRRPFQR
Sbjct: 631 TAASKPNSTVPKGEAKPKRRKKVRRPFQR 659


>g15029846 RIKEN cDNA 5730551F12 gene [Mus musculus]
           Length = 703

 Score =  533 bits (1357), Expect = e-150
 Identities = 250/269 (92%), Positives = 258/269 (94%)

Query: 1   MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 60
           MAGVPEDELNPFHVLGVEATASD ELKKAYRQLAVMVHPDKNHHPRAEEAFK+LRAAWDI
Sbjct: 435 MAGVPEDELNPFHVLGVEATASDTELKKAYRQLAVMVHPDKNHHPRAEEAFKILRAAWDI 494

Query: 61  VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 120
           VSN E+RKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP
Sbjct: 495 VSNPERRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 554

Query: 121 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 180
           KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH
Sbjct: 555 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 614

Query: 181 RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAPGA 240
           RVPYHISFGSR+PGT GRQRATP++PP DLQDFLSRIFQVPPG M NGNFFAAP P PG
```

```
Sbjct: 615 RVPYHISFGSRVPGTSGRQRATPESPPVDLQDFLSRIFQVPPGPMSNGNFFAAPHPGPGT 674

Query: 241 AAASKPNSTVPKGEAKPKRRKKVRRPFQR 269
            + S+PNS+VPKGEAKPKRRKKVRRPFQR
Sbjct: 675 TSTSRPNSSVPKGEAKPKRRKKVRRPFQR 703


>g26349793 unnamed protein product [Mus musculus]
           Length = 703

 Score =  531 bits (1354), Expect = e-150
 Identities = 250/269 (92%), Positives = 258/269 (94%)

Query: 1   MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 60
           MAGVPEDELNPFHVLGVEATASD ELKKAYRQLAVMVHPDKNHHPRAEEAFK+LRAAWDI
Sbjct: 435 MAGVPEDELNPFHVLGVEATASDTELKKAYRQLAVMVHPDKNHHPRAEEAFKILRAAWDI 494

Query: 61  VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 120
           VSN E+RKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP
Sbjct: 495 VSNPERRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 554

Query: 121 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 180
           KSA YCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH
Sbjct: 555 KSAGYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 614

Query: 181 RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAPGA 240
           RVPYHISFGSR+PGT GRQRATP++PPADLQDFLSRIFQVPPG M NGNFFAAP P PG
Sbjct: 615 RVPYHISFGSRVPGTSGRQRATPESPPADLQDFLSRIFQVPPGPMSNGNFFAAPHPGPGT 674

Query: 241 AAASKPNSTVPKGEAKPKRRKKVRRPFQR 269
            + S+PNS+VPKGEAKPKRRKKVRRPFQR
Sbjct: 675 TSTSRPNSSVPKGEAKPKRRKKVRRPFQR 703


>g26337271 unnamed protein product [Mus musculus]
           Length = 678

 Score =  454 bits (1156), Expect = e-126
 Identities = 216/238 (90%), Positives = 221/238 (92%)

Query: 1   MAGVPEDELNPFHVLGVEATASDVELKKAYRQLAVMVHPDKNHHPRAEEAFKVLRAAWDI 60
           MAGVPEDELNPFHVLGVEATASD ELKKAYRQLAVMVHPDKNHHPRAEEAFK+LRAAWDI
Sbjct: 435 MAGVPEDELNPFHVLGVEATASDTELKKAYRQLAVMVHPDKNHHPRAEEAFKILRAAWDI 494

Query: 61  VSNAEKRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 120
           VSN E+RKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP
Sbjct: 495 VSNPERRKEYEMKRMAENELSRSVNEFLSKLQDDLKEAMNTMMCSRCQGKHRRFEMDREP 554

Query: 121 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 180
           KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH
Sbjct: 555 KSARYCAECNRLHPAEEGDFWAESSMLGLKITYFALMDGKVYDITEWAGCQRVGISPDTH 614

Query: 181 RVPYHISFGSRIPGTRGRQRATPDAPPADLQDFLSRIFQVPPGQMPNGNFFAAPQPAP 238
           RVPYHISFGSR+PGT GRQRATP++PPADLQDFLSRIFQVP G            P P
Sbjct: 615 RVPYHISFGSRVPGTSGRQRATPESPPADLQDFLSRIFQVPSGADVQWELLCRTSPWP 672


  Database: genpept137
    Posted date:  Sep 11, 2003 11:22 AM
  Number of letters in database: 474,463,515
  Number of sequences in database:  1,534,369

Lambda     K      H
   0.318    0.133    0.404

Gapped
Lambda     K      H
```

```
      0.270   0.0470    0.230


   Matrix: BLOSUM62
   Gap Penalties: Existence: 11, Extension: 1
   Number of Hits to DB: 262275488
   Number of Sequences: 1534369
   Number of extensions: 10885493
   Number of successful extensions: 48330
   Number of sequences better than 10.0: 1367
   Number of HSP's better than 10.0 without gapping: 1113
   Number of HSP's successfully gapped in prelim test: 254
   Number of HSP's that attempted gapping in prelim test: 46724
   Number of HSP's gapped (non-prelim): 1477
   length of query: 269
   length of database: 474,463,515
   effective HSP length: 57
   effective length of query: 212
   effective length of database: 387,004,482
   effective search space: 82044950184
   effective search space used: 82044950184
   T: 11
   A: 40
   X1: 16 ( 7.3 bits)
   X2: 38 (14.8 bits)
   X3: 64 (24.9 bits)
   S1: 41 (21.7 bits)
```

Graphical Viewer

Submit sequences to: BLAST2 Submit Reset

IncyteGenomics

Exhibit A of Iyer Declaration
with Response dated 4/23/04
In USSN: 10/049,742

Vishwanath R. Iyer

Assistant Professor
Section of Molecular Genetics and Microbiology
Institute of Cellular and Molecular Biology
MBB 3.212A, University of Texas at Austin
Austin, TX 78712-0159
Phone: 512-232-7833
Fax: 512-232-3432
Email: vishy@mail.utexas.edu

Education/Training

Bombay University, Mumbai, India	B.Sc. (1987), Chemistry & Biochemistry
M. S. University of Baroda, Baroda, India	M.Sc. (1989), Biotechnology
Harvard University, Cambridge MA	Ph.D. (1996), Genetics
Stanford University, Stanford CA	Post-doctoral (1996-2000), Genomics

Research Experience

9/00-5/03 Assistant professor, Section of Molecular Genetics and Microbiology, University of Texas, Austin TX

- Global transcriptional control in yeast
- Gene expression programs during human cell proliferation
- Genome-wide transcription factor targets in yeast and human
- Collaborative microarray facility

5/96-8/00 Post-doctoral fellow Stanford University, Stanford CA
(Advisor: Dr. Patrick O. Brown)

- Yeast whole-genome ORF and intergenic microarrays
- Human cDNA microarrays for expression profiling

9/89-4/96 Graduate student Harvard University, Cambridge MA
(Advisor: Dr. Kevin Struhl)

- Yeast transcriptional regulation

Honours and Awards

Government of India Biotechnology Fellowship (1987-1989)
University Grants Commission Junior Research Fellowship (1989)
Stanford University/NHGRI Genome Training Grant (1996)

Invited Conference talks (selected)

Invited Lecturer, NEC-Princeton Lectures in Biophysics
Princeton, NJ (June 1998)
Plenary Session Speaker, HGM '99 (HUGO Human Genome Meeting)
Brisbane, Australia (April 1999)
Invited Speaker, Gordon Research Conference "Human Molecular Genetics"
Newport, RI (August 2001)

Invited Speaker, Nature Genetics "Oncogenomics 2002" Conference
Dublin, Ireland (May 2002)

Invited Speaker, "Pathology Bioinformatics" Symposium, University of Michigan, Ann Arbor, MI (November 2002)

Invited Speaker, "Systems Biology: Genomic Approaches to Transcriptional Regulation" Cold Spring Harbor Laboratory Meeting (March 2003)

Symposium co-Chair and Speaker "Functional Genomics" American Society for Biochemistry and Molecular Biology Meeting, San Diego, CA (April 2003)

Invited Speaker in Functional Genomics (Gene Networks) Symposium, International Congress of Genetics, Melbourne Australia July 6-11 2003

Invited Speaker "BioArrays Europe 2003"
Cambridge, UK (Sep/Oct 2003)

Departmental Seminars

Texas A&M University Genetics and Biochemistry & Biophysics Departments, October 24 2002

New York University School of Medicine, Department of Biochemistry, November 20 2002

UT Southwestern Medical Center, Human Genetics Seminar Series, May 5 2002

UCLA School of Medicine, Department of Human Genetics
June 2 2003

National Human Genome Research Institute
June 12 2003

Sanger Institute of the Wellcome Trust, Hinxton, UK
Sep 2003

Other Professional Activities

Reviewer for *Genome Biology, Genome Research, Nature Genetics, Science* (1998-2003)

Instructor, Cold Spring Harbor Summer Course "Making and using DNA Microarrays" (2000 - 2003)

Member, NIDDK Special Emphasis Review Panel ZDK1 (2001-2002)

Publications

1. Iyer V. & Struhl, K. (1995) Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure, *EMBO J.* 14: 2570-2579.

2. Iyer V. & Struhl, K. (1995) Mechanism of differential utilization of the his3 TR and TC TATA elements, *Mol. Cell. Biol.* 15: 7059-7066.

3. Iyer V. & Struhl K. (1996) Absolute mRNA levels and transcription initiation rates in *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci*. (*USA*) 93:5208-5212.

4. DeRisi J. L., Iyer V. R. & Brown P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686

5. Marton M. J., DeRisi J. L., Bennett H. A., Iyer V. R., Meyer M. R., Roberts C. J., Stoughton R., Burchard J., Slade D., Dai H., Bassett D. E. Jr., Hartwell L. H., Brown P. O. & Friend S. H. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* 4:1293-1301

6. Lutfiyya L. L., Iyer V. R., DeRisi J., DeVit M. J., Brown P. O. & Johnston M. (1998) Characterization of three related glucose repressors and genes they regulate in *Saccharomyces cerevisiae. Genetics* 150:1377-1391

7. Spellman P. T., Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B., Brown P. O., Botstein D. & Futcher B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9:3273-3297

8. Iyer V. R., Eisen M. B., Ross D. T., Schuler G., Moore T., Lee J. C.,F., Trent J. M., Staudt L. M., Hudson Jr. J., Boguski M. S., Lashkari D., Shalon D., Botstein D. & Brown P. O. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* 283:83-87

9. DeRisi J. L. & Iyer V. R. (1999) Genomics and array technology. *Curr. Opin. Oncol.* 11:76-79

10. Ross D. T., Scherf U., Eisen M. B., Perou C. M., Spellman P., Iyer V. R., Rees C., Jeffrey S. S., Van de Rijn M., Waltham M., Pergamenschikov A., Lee J. C. F., Lashkari D., Shalon D., Myers T. G., Weinstein J. N., Botstein D., & Brown P. O. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* 24: 227-235

11. Sudarsanam P., Iyer V. R., Brown P. O. & Winston F. (2000) Whole-genome expression analysis of *snf/swi* mutants of *S. cerevisiae. Proc. Natl. Acad. Sci .(USA)* 97: 3364-3369

12. Tran H. G., Steger D. J., Iyer V. R., & Johnson A. D. (2000) The chromo domain protein Chd1p from budding yeast is an ATP-dependent chromatin-modifying factor *EMBO J* 19: 2323-2331

13. Gross C., Kelleher M., Iyer V. R., Brown P. O., & Winge D. R.. (2000) Identification of the copper regulon in *Saccharomyces cerevisiae* by DNA microarrays. *J. Biol. Chem.* 275: 32310-32316

14. Reid J. L., Iyer V. R., Brown P. O. & Struhl K. (2000) Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol. Cell* 6: 1297–1307

15. Iyer V. R., Horak C., Scafe C. S., Botstein D., Snyder M. & Brown P. O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF *Nature* 409: 533-538

16. Miki R., Kadota K., Bono H., Mizuno Y., Tomaru Y., Carninci P., Itoh M., Shibata K., Kawai J., Konno H., Watanabe S., Sato K., Tokusumi Y., Kikuchi N., Ishii Y., Hamaguchi Y., Nishizuka I., Goto H., Nitanda H., Satomi S., Yoshiki A., Kusakabe M., DeRisi J.L., Eisen M.B., Iyer V.R., Brown P.O., Muramatsu M., Shimada H., Okazaki Y. & Hayashizaki Y. (2001) Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays *Proc. Natl. Acad. Sci.* (*USA*) 98: 2199-2204

17. Pollack J. R. & Iyer V.R. (2002) Characterizing the physical genome. *Nature Genetics* 32 suppl: 515-521

18. Iyer V. R. Microarray-based detection of DNA protein interactions: Chromatin Immunoprecipitation on Microarrays, in *DNA Microarrays: A Molecular Cloning Manual* (eds. Bowtell, D. & Sambrook, J.) 453-463 (Cold Spring Harbor Laboratory Press, 2003).
*(*not peer reviewed*)

19. Killion, P., Sherlock G. and Iyer V. R. (2003) The Longhorn Array Database, an open-source implementation of the Stanford Microarray Database *BMC Bioinformatics* 4: 32

20. Hahn J. S., Hu Z., Thiele D. J. & Iyer V. R. Genome-Wide Analysis of the Biology of Stress Responses Through Heat Shock Transcription Factor (submitted to *PNAS*)

21. Kim J. & Iyer V.R. The global role of TBP recruitment to promoters in mediating gene expression profiles (manuscript in preparation)

Current/Pending Research Support

U01 AA13518-01 Adron Harris (PI) 25% effort
9/28/01 - 9/27/06
NIH/NIAAA
"INIA: Microarray Core"
This proposal was a response to the Integrative Neuroscience Initiative on Alcoholism (INIA) RFA-AA-01-002. The overall goal is to support the use of microarray technology to define changes in gene expression that either predict or accompany excessive alcohol consumption.
Role: Co-investigator

003658-0223-2001 Iyer (PI) 16% effort
01/01/02 - 08/31/04
Texas Higher Education Coordinating Board (ARP)
"Microarray based global mapping of DNA-protein interactions at promoters in human cells"
This is a pilot project to map the in vivo interactions of transcription factors with human promoters
Role: PI

Information Technology Research 0325116 R. Mooney (PI) 9% effort
09/01/03 - 08/31/07
NSF
"Feedback from Multi-Source Data Mining to Experimentation for Gene Network Discovery"
Role: Co-investigator

1 R01 CA95548-01A2 (pending) Iyer (PI) 25% effort
12/1/03 – 11/30/08
NIH
"Analysis of genome-wide transcriptional control in yeast"
This is a project to identify stress responsive transcription factor targets in yeast through the use of DNA microarrays
Role: PI

Breast Cancer Idea Award (pending) Iyer (PI) 10% effort
1/1/04 – 12/31/06
US Army Medical Research and Materiel Command
"Genome-wide chromosomal targets of oncogenic transcription factors"
This is a project aimed at identifying direct chromosomal targets of c-myc and ER in human cells through the use of a novel sequence tag analysis method.
Role: PI
003658-0531-2003 (pending) Marcotte (PI) 8% effort
01/01/04 - 12/31/05
Texas Higher Education Coordinating Board (ATP)
"Cell arrays: A novel high–throughput platform for measuring gene function on a genomic scale"
This proposal is aimed at developing a novel microarray based platform for automated, high–throughput microscopic imaging of cells, allowing rapid and systematic evaluation of gene function.

Exhibit B of Iyer Declaration
with Response dated 4/23/04
In USSN: 10/049,742

Fischer-Vize, *Science* 270, 1828 (1995).

35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* 6, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* 88, 263 (1991); B. Tschiersch *et al.*, *EMBO J.* 13, 3822 (1994); M. T. Madireddi *et al.*, *Cell* 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* 93, 7137 (1996).
36. P. M. Palosaari *et al.*, *J. Biol. Chem.* 266, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E. Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* 58, 4068 (1992); V. Sharma, K. Suvarna, R. Meganathan, M. E. Hudspeth, *J. Bacteriol.* 174, 5057 (1992); M. Kanazawa *et al.*, *Enzyme Protein* 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* 178, 3015 (1996).
37. M. Ho *et al.*, *Cell* 77, 869 (1994).
38. W. Hendriks *et al.*, *J. Cell Biochem.* 59, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with sequence analysis; Lawrence Livermore National Laboratory for the flow-sorted Y cosmid library; and P. Bain, A. Bortvin, A. de la Chapelle, G. Fink, K. Jegalian, T. Kawaguchi, E. Lander, H. Lodish, P. Matsudaira, D. Menke, U. RajBhandary, R. Reijo, S. Rozen, A. Schwartz, C. Sun, and C. Tilford for comments on the manuscript. Supported by NIH.

28 April 1997; accepted 9 September 1997

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (*1*, *2*), provide a practical and economical tool for studying gene expression on a very large scale (*3*–*6*).

Saccharomyces cerevisiae is an especially favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, *cis* regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (*7*). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (*8*). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by using a simple robotic printing device (*9*). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (*10*). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (*11*) and then hybridized to the microarrays (*12*). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (*13*).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (*14*). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305-5428, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

to any gene whose function is known (*15*). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (*ALD2*) and acetyl–coenzyme A(CoA) synthase (*ACS1*), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes *PCK1*, encoding phosphoenolpyruvate carboxykinase, and *FBP1*, encoding fructose 1,6-biphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c–related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (*13*). A noteworthy and illuminating exception was that the genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitchondrial biogenesis (*13*). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (*16–20*). A search in the promoter regions of the remaining two genes, *ACR1* and *IDP2*, revealed that *ACR1*, a gene essential for *ACS1* activity, also possessed a consensus CSRE motif, but interestingly, *IDP2* did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception

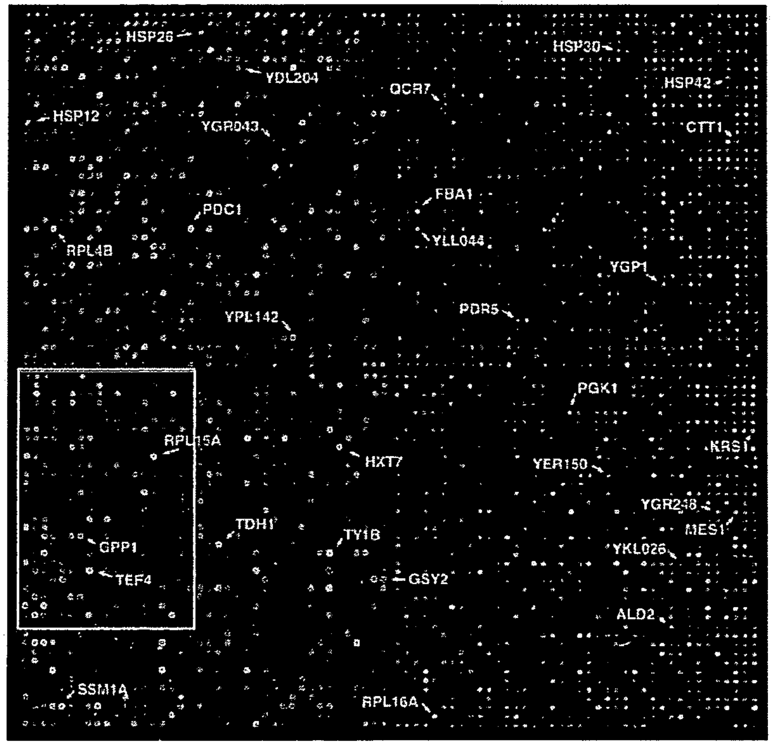


Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (*9*). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (*49*). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of $<5 \times 10^6$ cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of $\sim 2 \times 10^8$ cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP–labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP–labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of *HSP42*, have previously been shown to be controlled at least in part by these elements (*21–24*). Inspection of the sequences upstream of *HSP42* and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including *HSP30*, *ALD2*, *OM45*, and 10 uncharacterized ORFs (*25*)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex *HAP2,3,4* has been shown to be responsible for induction of several genes important for respiration (*26–28*). This complex binds a degenerate consensus sequence known as the CCAAT box (*26*). Computer analysis, using the consensus sequence TNRYTGGB (*29*), has suggested that a large number of genes involved in respiration may be specific targets of *HAP2,3,4* (*30*). Indeed, a putative *HAP2,3,4* binding site could be found in the sequences upstream of each of the seven cytochrome c–related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome c–related genes that were induced, *HAP2,3,4* binding sites were present in all but one. Significantly, we found that transcription of *HAP4* itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS$_{rpg}$) that is recognized by the Rap1 DNA-binding protein (*31*, *32*). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (*33*). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (*34*). Indeed, we observed that the abundance of *RAP1* mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, *HAP4* and *SIP4*, were induced by a factor of more than threefold at the diauxic shift. *SIP4* encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the "master regulator" of glucose repression (*35*). The eightfold induction of *SIP4* upon depletion of glucose strongly suggests a role in the induction of downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (*36*). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expression ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (*37*, *38*). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

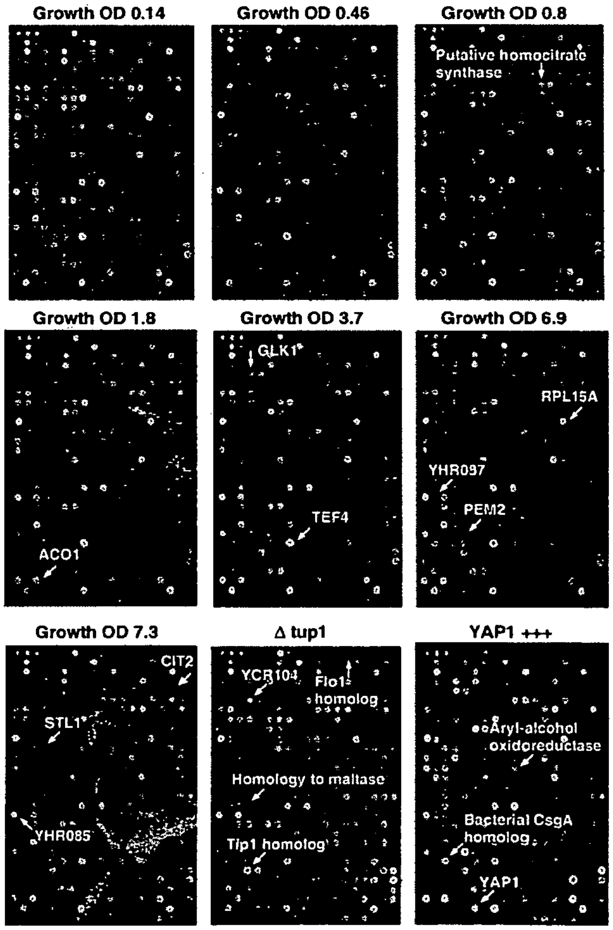


Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1*Δ mutation and *YAP1* overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (*13*). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.

by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (*39*). Tup1 has also been implicated in repression of oxygen-regulated, mating-type–specific, and DNA-damage–inducible genes (*40*).

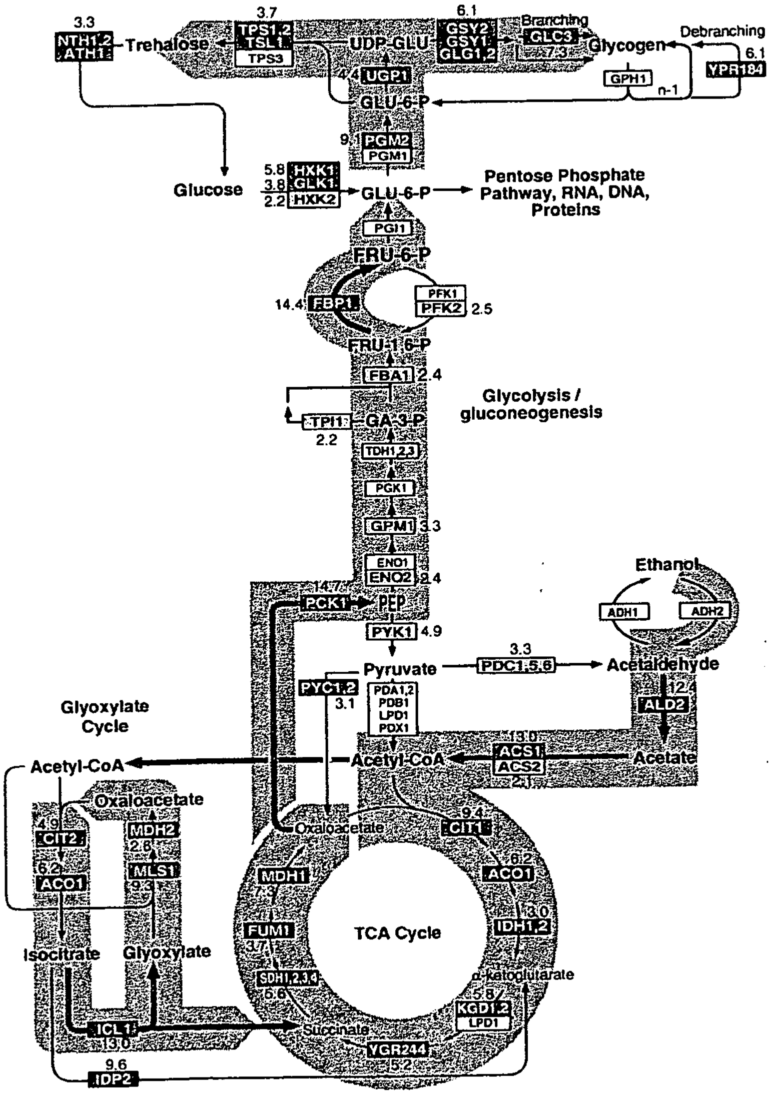


Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs YGR244 and YPR184 show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1*Δ) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (*11*). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1*Δ strain, and thus presumably repressed by Tup1 (*41*). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1*Δ mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (*13*)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α-glucosidases, the mating-type–specific genes *MFA1* and *MFA2*, and the DNA damage–inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1*Δ strain, providing a positive control in the experiment (*42*).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (*15*) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as Tip1 and Tir1/Srp1 which are induced by cold shock and other stresses (*43*), and similar, serine-poor proteins, the seripauperins (*44*). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1*Δ strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a *MAT* α strain in which *MFA1* and *MFA2*, the genes encoding the a-factor mating pheromone precursor, are normally repressed. In the isogenic *tup1*Δ strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a *MAT*A strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, *o*-phenanthroline, heavy metals, and osmotic stress (*45*). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GAL1-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

Of the 17 genes whose mRNA levels increased by more than threefold when *YAP1* was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (*46, 47*). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing Yap1. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for Yap1-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (*48*). About two-thirds of the genes that were induced by more than threefold upon Yap1 overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by Yap1. The absence of canonical Yap1-bind-

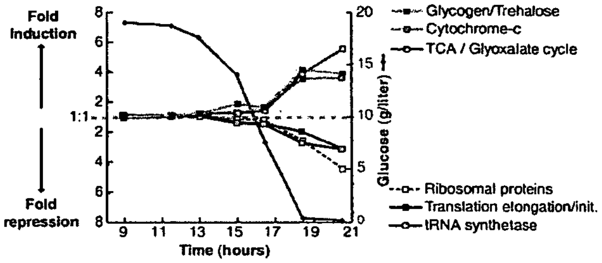


Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.

Table 1. Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (*50*). Positions of the canonical Yap1 binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of Yap1 site from ATG	Gene	Description	Fold-increase
YNL331C			Putative aryl-alcohol reductase	12.9
YKL071W	162–222 (5 sites)		Similarity to bacterial csgA protein	10.4
YML007W		YAP1	Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W			Putative aryl-alcohol reductase	6.5
YML116W	409	ATR1	Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	OYE3	NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	OYE2	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to *A. thaliana* zeta-crystallin homolog	3.7
YOL126C		MDH2	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and networks. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

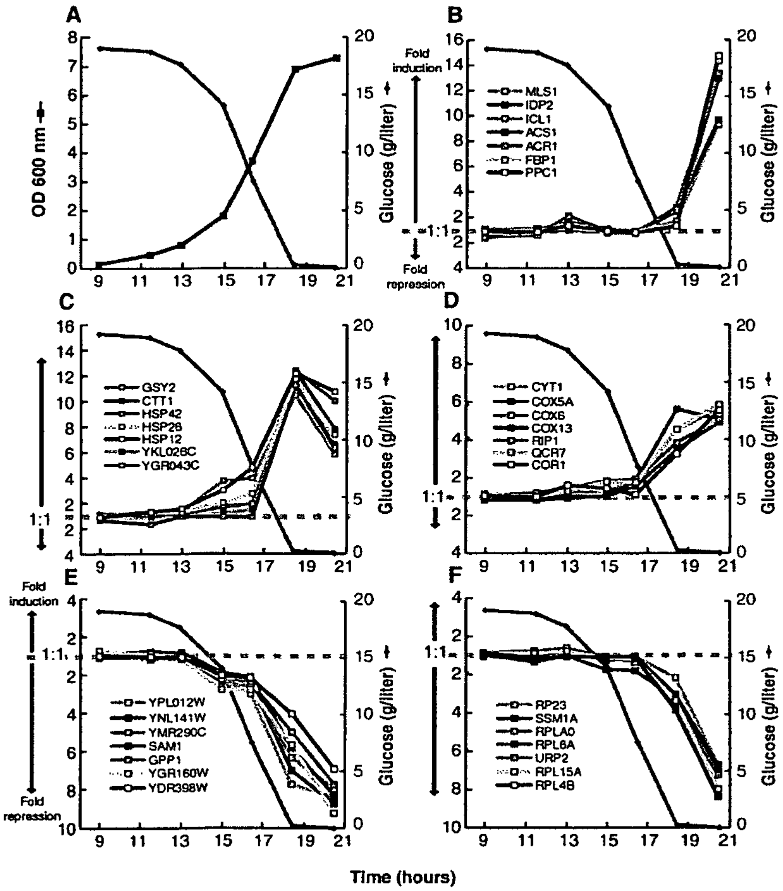


Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (**A**) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (**B**) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (**C**) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contain STRE motif repeats in their upstream promoter regions. (**D**) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (**E**) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (**F**) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* 270, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* 6, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi *et al.*, *Nature Genet.* 14, 457 (1996).
5. D. J. Lockhart *et al.*, *Nature Biotechnol.* 14, 1675 (1996).
6. M. Chee *et al.*, *Science* 274, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast* Saccharomyces: *Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100-μl PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3× standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-

tion, the bound DNA was denatured by a 2-min incubation in distilled water at ∼95°C. The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.

10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251) Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at −80°C.
11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 μg of polyadenylated [poly(A)$^+$] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 μM for dATP, dCTP, and dGTP and 200 μM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 μM. The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with of 470 μl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ∼5 μl, using Centricon-30 microconcentrators (Amicon).
12. Purified, labeled cDNA was resuspended in 11 μl of 3.5× SSC containing 10 μg poly(dA) and 0.3 μl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ∼8 to 12 hours in a water bath at 62°C. Before scanning, slides were washed in 2× SSC, 0.2% SDS for 5 min, and then 0.05× SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.
13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html
14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.
15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: *Saccharomyces* Genome Database (genome-www.stanford.edu), Yeast Protein Database (quest7.proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/mips/yeast/index.htmlx).
16. A. Scholer and H. J. Schuller, *Mol. Cell. Biol.* **14**, 3613 (1994).
17. S. Kratzer and H. J. Schuller, *Gene* **161**, 75 (1995).
18. R. J. Haselbeck and H. L. McAlister, *J. Biol. Chem.* **268**, 12116 (1993).
19. M. Fernandez, E. Fernandez, R. Rodicio, *Mol. Gen. Genet.* **242**, 727 (1994).
20. A. Hartig *et al.*, *Nucleic Acids Res.* **20**, 5677 (1992).
21. P. M. Martinez *et al.*, *EMBO J.* **15**, 2227 (1996).
22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, *Mol. Cell. Biol.* **15**, 6232 (1995).
23. H. Ruis and C. Schuller, *Bioessays* **17**, 959 (1995).
24. J. L. Parrou, M. A. Teste, J. Francois, *Microbiology* **143**, 1891 (1997).
25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.
26. S. L. Forsburg and L. Guarente, *Genes Dev.* **3**, 1166 (1989).
27. J. T. Olesen and L. Guarente, *ibid.* **4**, 1714 (1990).
28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, *Mol. Microbiol.* **13**, 119 (1994).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B–C, G, or T; N–G, A, T, or C; R–A or G; and Y–C or T.
30. C. Fondrat and A. Kalogeropoulos, *Comput. Appl. Biosci.* **12**, 363 (1996).
31. D. Shore, *Trends Genet.* **10**, 408 (1994).
32. R. J. Planta and H. A. Raue, *ibid.* **4**, 64 (1988).
33. The degenerate consensus sequence VYCYRNNCMNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (*30*), is WACAYCCRTACATYW, with up to three differences allowed.
34. S. F. Neuman, S. Bhattacharya, J. R. Broach, *Mol. Cell. Biol.* **15**, 3187 (1995).
35. P. Lesage, X. Yang, M. Carlson, *ibid.* **16**, 1921 (1996).
36. For example, we observed large inductions of the genes coding for *PCK1*, *FBP1* [Z. Yin *et al.*, *Mol. Microbiol.* **20**, 751 (1996)], the central glyoxylate cycle gene *ICL1* [A. Scholer and H. J. Schuller, *Curr. Genet.* **23**, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, *ACS1* [M. A. van den Berg *et al.*, *J. Biol. Chem.* **271**, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes *PYK1* and *PFK2* [P. A. Moore *et al.*, *Mol. Cell. Biol.* **11**, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase T *CTT1* [P. H. Bissinger *et al.*, *ibid.* **9**, 1309 (1989)] and several genes encoding small heat-shock proteins, such as *HSP12*, *HSP26*, and *HSP42* [I. Farkas *et al.*, *J. Biol. Chem.* **266**, 15602 (1991); U. M. Praekelt and P. A. Meacock, *Mol. Gen. Genet.* **223**, 97 (1990); D. Wotton *et al.*, *J. Biol. Chem.* **271**, 2717 (1996)].
37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, *FBP1* and *PCK1*) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (*46*).
38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, *ADH1* and *ADH2*, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as *HXK1/HXK2* (77% identical) [P. Herrero *et al.*, *Yeast* **11**, 137 (1995)], *MLS1/DAL7* (73% identical) (*20*), and *PGM1/PGM2* (72% identical) [D. Oh, J. E. Hopper, *Mol. Cell. Biol.* **10**, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.
39. F. E. Williams, U. Varanasi, R. J. Trumbly, *Mol. Cell. Biol.* **11**, 3307 (1991).
40. D. Tzamarias and K. Struhl, *Nature* **369**, 758 (1994).
41. Differences in mRNA levels between the tup1Δ and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the tup1Δ strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.
42. The tup1Δ mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of *TUP1* and an Eco R I site 124 base pairs before the stop codon of the *TUP1* gene.
43. L. R. Kowalski, K. Kondo, M. Inouye, *Mol. Microbiol.* **15**, 341 (1995).
44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, *Gene* **148**, 149 (1994).
45. D. Hirata, K. Yano, T. Miyakawa, *Mol. Gen. Genet.* **242**, 250 (1994).
46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, *Appl. Environ. Microbiol.* **60**, 1783 (1994).
47. A. Muheim *et al.*, *Eur. J. Biochem.* **195**, 369 (1991).
48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, *J. Biol. Chem.* **269**, 32592 (1994).
49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm.stanford.edu/pbrown. Images were scanned at a resolution of 20 μm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.
50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold)

51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the Tup1 deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997

Exhibit C of Iyer Declaration
with Response dated 4/23/04
In USSN: 10/049,742

Drug target validation and identification of secondary drug target effects using DNA microarrays

MATTHEW J. MARTON[1], JOSEPH L. DERISI[2], HOLLY A. BENNETT[1], VISHWANATH R. IYER[2],
MICHAEL R. MEYER[1], CHRISTOPHER J. ROBERTS[1], ROLAND STOUGHTON[1], JULJA BURCHARD[1],
DAVID SLADE[1], HONGYUE DAI[1], DOUGLAS E. BASSETT, JR[1]., LELAND H. HARTWELL[3],
PATRICK O. BROWN[2] & STEPHEN H. FRIEND[1]

[1]Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034, USA
[2]Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute
Stanford, California 94305-5428, USA
[3]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, Washington 98109, USA
Correspondence should be addressed to S.H.F; email: sfriend@rosetta.org

We describe here a method for drug target validation and identification of secondary drug target effects based on genome-wide gene expression patterns. The method is demonstrated by several experiments, including treatment of yeast mutant strains defective in calcineurin, immunophilins or other genes with the immunosuppressants cyclosporin A or FK506. Presence or absence of the characteristic drug 'signature' pattern of altered gene expression in drug-treated cells with a mutation in the gene encoding a putative target established whether that target was required to generate the drug signature. Drug dependent effects were seen in 'targetless' cells, showing that FK506 affects additional pathways independent of calcineurin and the immunophilins. The described method permits the direct confirmation of drug targets and recognition of drug-dependent changes in gene expression that are modulated through pathways distinct from the drug's intended target. Such a method may prove useful in improving the efficiency of drug development programs.

Good drugs are potent and specific; that is, they must have strong effects on a specific biological pathway and minimal effects on all other pathways. Confirmation that a compound inhibits the intended target (drug target validation) and the identification of undesirable secondary effects are among the main challenges in developing new drugs. Comprehensive methods that enable researchers to determine which genes or activities are affected by a given drug might improve the efficiency of the drug discovery process by quickly identifying potential protein targets, or by accelerating the identification of compounds likely to be toxic. DNA microarray technology, which permits simultaneous measurement of the expression levels of thousands of genes, provides a comprehensive framework to determine how a compound affects cellular metabolism and regulation on a genomic scale[1-11]. DNA microarrays that contain essentially every open reading frame (ORF) in the *Saccharomyces cerevisiae* genome have already been used successfully to explore the changes in gene expression that accompany large changes in cellular metabolism or cell cycle progression[7-10].

In the modern drug discovery paradigm, which typically begins with the selection of a single molecular target, the ideal inhibitory drug is one that inhibits a single gene product so completely and so specifically that it is as if the gene product were absent. Treating cells with such a drug should induce changes in gene expression very similar to those resulting from deleting the gene encoding the drug's target. Here we have compared the genome-wide effects on gene expression that result from deletions of various genes in the budding yeast *S. cerevisiae* to the effects on gene expression that result from treatment with known inhibitors of those gene products. Using the calcineurin signaling pathway as a model system, we tested an approach that permits identification of genes that encode proteins specifically involved in pathways affected by a drug. The FK506 characteristic pattern, or 'signature', of altered gene expression was not observed in mutant cells lacking proteins inhibited by FK506 (for example, a calcineurin or FK506-binding-protein mutant strain), but was observed in mutants deleted for genes in pathways unrelated to FK506 action (for example, a cyclophilin mutant strain). Conversely, the cyclosporin A (CsA) signature was not observed in CsA-treated calcineurin or cyclophilin mutant strains, but was seen in an FK506-binding-protein mutant strain treated with CsA. The method also demonstrates that FK506, a clinically used immunosuppressant, has 'off-target' effects that are independent of its binding to immunophilins. Thus, the approach we describe may provide a way to identify the pathways altered by a drug and to detect drug effects mediated through unintended targets.

Null mutants phenocopy drug-treated cells on a genomic scale

To test whether a null mutation in a drug target serves as a model of an ideal inhibitory drug, we examined the effects on gene expression associated with pharmacological or genetic inhibition of calcineurin function. Calcineurin is a highly conserved calcium- and calmodulin-activated serine/threonine protein phosphatase implicated in diverse processes dependent on calcium signaling[12-13]. In budding yeast, calcineurin is required for intracellular ion homeostasis[14], for adaptation to prolonged mating pheromone treatment[15] and in the regulation of

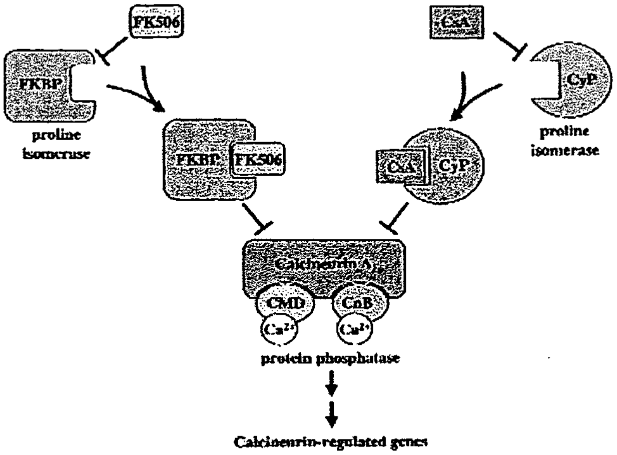


Fig. 1 Model of antagonism of the calcineurin signaling pathway mediated by FK506 and cyclosporin A (CsA). Calcineurin activity is composed of a catalytic subunit (calcineurin A, encoded in yeast by the *CNA1* and *CNA2* genes), and calcium-binding regulatory subunits calmodulin (CMD) and calcineurin B (CnB). After entering cells, FK506 and CsA specifically bind and inhibit the peptidyl-proline isomerase activity of their respective immunophilins, FK506 binding proteins (FKBP) and cyclophilins (CyP). The most abundant immunophilins in yeast (Fpr1 and Cph1) are thought to mediate calcineurin inhibition. Drug–immunophilin complexes bind and inhibit the calcium- and calmodulin-stimulated phosphatase calcineurin. Among the substrates of calcineurin are transcriptional activators that act to modulate gene expression.

the onset of mitosis[16]. In mammals, calcineurin has been implicated in T-cell activation[12], in apoptosis[17], in cardiac hypertrophy[18] and in the transition from short-term to long-term memory[19]. In both organisms, calcineurin activity is inhibited by FK506 and CsA, immunosuppressant drugs whose effects on calcineurin are mediated through families of intracellular receptor proteins called immunophilins[12,20] (Fig. 1). To assess the effects of pharmacologic inhibition of calcineurin, wild-type *S. cerevisiae* was grown to early logarithmic phase in the presence or absence of FK506 or CsA. Isogenic cells, from which the genes encoding the catalytic subunits of calcineurin (*CNA1* and *CNA2*) had been deleted[21] (referred to as the *cna* or calcineurin mutant), were grown in parallel, in the absence of the drug. Fluorescently-labeled cDNA was prepared by reverse transcription of polyA$^+$ RNA in the presence of Cy3- or Cy5-deoxynucleotide triphosphates and then hybridized to a microarray containing more than 6,000 DNA probes representing 97% of the known or predicted ORFs in the yeast genome. Simultaneous hybridization of Cy5-labeled cDNA from mock-treated cells and Cy3-labeled cDNA from cells treated with 1 µg/ml FK506 allowed the effect of drug treatment on mRNA levels of each ORF to be determined (Fig. 2*a* and *b* and data not shown). Similarly, effects of the calcineurin mutations on the mRNA levels of each gene were assessed by simultaneous hybridization of Cy5-labeled cDNA from wild-type cells and Cy3-labeled cDNA from the calcineurin mutant strain(Fig. 2*c*). For each comparison of this kind, reported expression ratios are the average of at least two hybridizations in which the Cy3 and Cy5 fluors were reversed to remove biases that may be introduced by gene-specific differences in incorporation of the two fluors (data not shown).

Treatment with FK506 in these growth conditions resulted in a signature pattern of altered gene expression in which mRNA levels of 36 ORFs changed by more than twofold (http://www.rosetta.org). A very similar pattern of altered gene expression was observed when the calcineurin mutant strain was compared to wild-type cells. Comparison of the changes in mRNA expression of each gene resulting from treatment of wild-type cells with FK506 with mRNA expression changes resulting from deletion of the calcineurin genes showed the considerable similarity of the global transcript alterations in response to the two perturbations (Fig. 2*b*–*d*). Quantification of this similarity using the correlation coefficient (ρ) showed large correlations between the FK506 treatment signature and the calcineurin deletion signature (ρ = 0.75 ± 0.03), as well as the CsA treatment signature (ρ = 0.94±0.02), but not with a randomly selected deletion mutant strain (deleted for the *YER071C* gene; ρ = -0.07 ± 0.04; Fig. 2*e*). The FK506 treatment signature was also compared with those of more than 40 other deletion mutant strains or drug-treatments thought to affect unrelated pathways, and none had statistically significant correlations. These data establish that genetic disruption of calcineurin function provides a close and specific phenocopy of treatment with FK506 or CsA.

To avoid generalizing from a single example, we also compared the effects of treatment of wild-type cells with 3-aminotriazole (3-AT) with the effects of deletion of the *HIS3* gene. *HIS3* encodes imidazoleglycerol phosphate dehydratase, which catalyzes the seventh step of the histidine biosynthetic pathway in yeast[22]; 3-AT is a competitive inhibitor of this enzyme that triggers a large transcriptional amino-acid starvation response[23]. Microarray analysis of wild-type and isogenic *his3*-deficient strains demonstrated the expected large genome-wide transcriptional responses (involving more than 1,000 ORFs) resulting from treatment with 3-AT (Fig. 3*a*) or from *HIS3* deletion (Fig. 3*c*). Quantitative comparison of the 3-AT treatment signature and the *his3* mutant signature showed a high level of correlation (*ρ*= 0.76 ± 0.02) that even extended to genes that experienced small changes in expression level (Fig. 3*b*). As a negative control, the correlations between the 3-AT treatment signature or the *his3* mutant signature and the calcineurin mutant strain were not statistically significant (*ρ* = 0.09 ± 0.06 and -0.01 ± 0.04, respectively). That both the calcineurin/FK506 and the *his3*/3-AT comparisons were highly correlated indicates that in many cases the expression profile resulting from a gene deletion closely resembles the expression profile of wild-type cells treated with an inhibitor of that gene's product.

'Decoder' strategy: Drug target validation with deletion mutants

Because pharmacological inhibition of different targets might give similar or identical expression profiles, simple comparison of drug signatures to mutant signatures is unlikely to unambiguously identify a drug's target. To overcome this limitation, an additional 'decoder' step is used. We first compare the expression profile of wild-type drug-treated cells to the expression profiles from a panel of genetic mutant strains, using a correlation coefficient metric. Mutant strains whose expression profile is similar to that of drug-treated wild-type cells are selected and subjected to drug treatment, generating the drug signature in the mutant strain (that is, the mutant drug signature). If the mutated gene encodes a protein involved in a pathway affected by the drug, we expect the drug signature in mutant cells to be different (or absent, for an ideal drug) from the drug signature seen in wild-type cells.

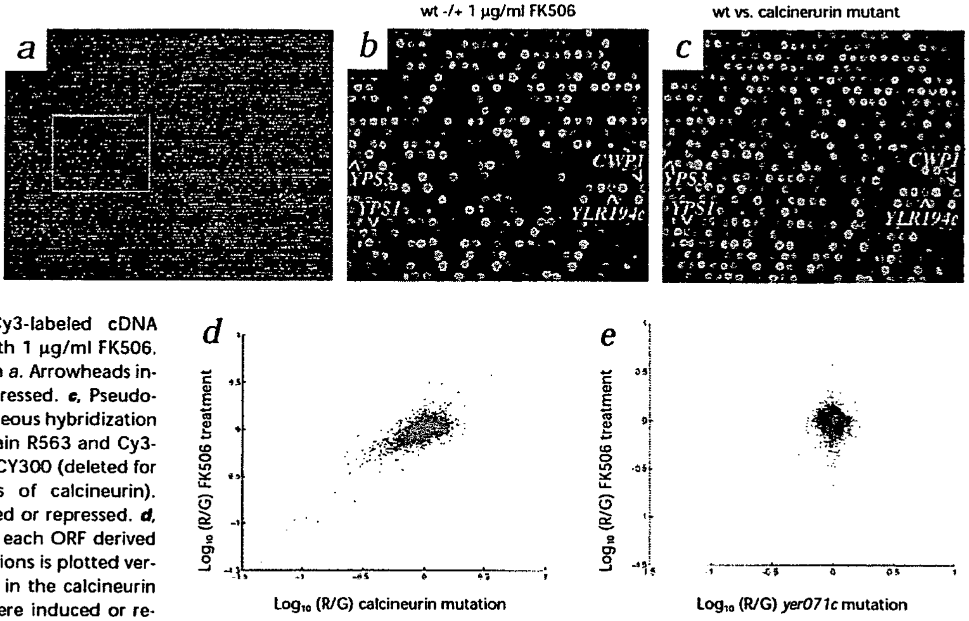


Fig. 2 Expression profiles from FK506-treated wild-type (wt) cells and a calcineurin-disruption mutant strain share a genome-wide correlation. DNA microarray analysis showing changes in gene expression resulting from FK506 treatment (*a* and *b*) or from genetic disruption of genes encoding calcineurin (*c*). ***a***, Pseudocolor image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from mock-treated strain R563 and Cy3-labeled cDNA (green) from strain R563 treated with 1 µg/ml FK506. ***b***, Enlarged view of the boxed area in *a*. Arrowheads indicate specific ORFs induced or repressed. ***c***, Pseudocolor image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from strain R563 and Cy3-labeled cDNA (green) from strain MCY300 (deleted for the *CNA1,CNA2* catalytic subunits of calcineurin). Arrows indicate specific ORFs induced or repressed. ***d***, The $\log_{10}$ of the expression ratio for each ORF derived from the FK506 treatment hybridizations is plotted versus the $\log_{10}$ of the expression ratio in the calcineurin mutant hybridizations. ORFs that were induced or repressed in both experiments are shown as green and red dots, respectively. ***e***, The $\log_{10}$ of the expression ratio for each ORF derived from the FK506 treatment hybridizations is plotted versus the $\log_{10}$ of the expression ratio in the *yer071c* mutant hybridizations. No ORFs were induced or repressed in both experiments.

To illustrate this, we treated the *his3* mutant strain with 3-AT. The signature pattern of altered gene expression resulting from treatment of the mutant strain with 3-AT was much less complex than that of the 3-AT signature in wild-type cells (Fig. 4). This is seen simply by examining plots of mean intensity of the hybridization signal (which approximately reflects level of expression) versus the expression ratio for each ORF (Fig. 4). Genes that were expressed at higher or lower levels in 3-AT treated cells or in *his3* mutant cells are shown as red and green dots, respectively. We analyzed the 3-AT signature in wild-type (Fig. 4a) and *his3* mutant cells (Fig. 4c), as well as the *his3* mutant strain signature (Fig. 4b). Whereas histidine limitation induced by 3-AT induced more than 1,000 transcription-level changes in the wild-type strain, few or no transcript level changes were induced by treatment of the *his3*-deletion strain with 3-AT. This indicates that with the growth conditions used, essentially all of the effects of 3-AT depend on or are mediated through the HIS3 gene product.

Applying this approach to the calcineurin signaling pathway showed the specificity of the method. The calcineurin mutant strain and strains with deletions in the genes encoding the most abundant immunophilins in yeast[12] (*CPH1* and *FPR1*) were treated with either FK506 or CsA to determine the profiles of altered gene expression resulting from drug treatment of the mutant cells (that is, mutant +/- drug). We compared the drug signatures in the mutants to the wild-type drug signature using the correlation coefficient metric (Table 1). Although the signature generated by treatment of wild-type cells with FK506 was highly correlated to the calcineurin mutant strain signature (ρ = 0.75 ± 0.03), it bore no similarity to the profile after treatment of the calcineurin mutant strain with FK506 (ρ = -0.01 ± 0.07). This indicates that FK506 was unable to elicit its normal transcriptional response in the calcineurin mutant strain. Likewise, treatment of the *fprl* mutant strain with FK506 elicited an expression profile that was not correlated to the FK506 signature in the wild-type strain (ρ = -0.23 ± 0.07), indicating that the *FPR1* gene product is likely to be involved in the pathway affected by FK506. The same was true for the *cna fprl* mutant strain. In contrast, treatment of the *cphl* mutant strain with FK506 generated an expression profile highly correlated with the wild-type FK506 expression profile (ρ = 0.79 ± 0.03), indicating the *cphl* mutation did not block the mode of action of FK506 and thus is not directly involved in the pathway affected by FK506. We tabulated the change in expression in response to FK506 in different mutant strains for all ORFs with expression ratios greater than 1.8 in FK506-treated cells or in the calcineurin mutant strain (Fig. 5*a*).The calcineurin mutant strain signature and the FK506 responses in wild-type and the *cphl* mutant strain are similar, and there are no transcript-level changes (seen in black) for treatment of the calcineurin, *fprl* and *cna fprl* mutant strains with FK506 (Fig. 5*a*).

Table 1 Signature correlation of expression ratios as a result of FK506 treatment in various mutant strains

	wild-type +/-FK506	*cna* +/-FK506	*fpr1* +/-FK506	*cna fpr1* +/-FK506	*cph1* +/-FK506
wild-type +/- FK506	0.93 ± 0.04	-0.01 ± 0.07	-0.23 ± 0.07	0.12 ± 0.07	0.79 ± 0.03

Signature correlation shows the absence of the FK506 signature specifically in the calcineurin (*cna*) and *fpr1* (major FK506 binding protein) deletion mutants. *cna* represents the mutant with deletions of the catalytic subunits of calcineurin, *CNA1* and *CNA2*. The correlation coefficient reported in the first column represents the correlation between two pairs of hybridizations from independent wild-type +/- FK506 experiments.

Similar experiments and analyses with CsA provided further validation of this approach. The expression profile elicited by treatment of wild-type cells with CsA was highly corre-

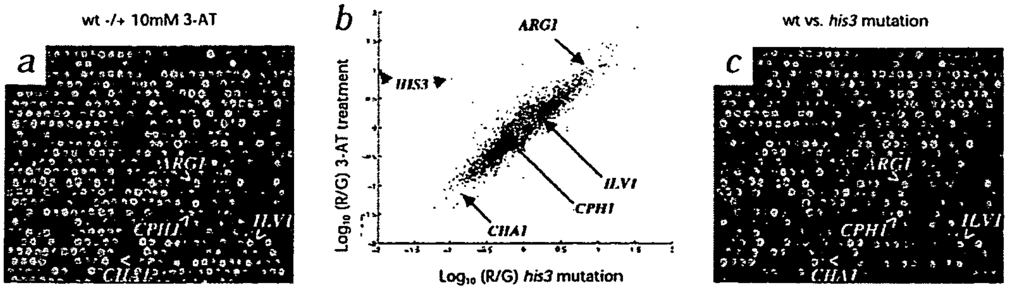


Fig. 3 Expression profiles from a *his3* mutant strain and wild-type (wt) cells treated with 3-AT share a genome-wide correlation. DNA microarray analysis showing changes in gene expression resulting from 3-AT treatment (*a*) or from genetic disruption of the *HIS3* gene (*c*). ***a***, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from mock-treated wild-type strain R491 and Cy3-labeled cDNA (green) from strain R491 treated with 10 mM 3-AT. ***b***, Plot of the $\log_{10}$ of the expression ratio for each ORF derived from the 3-AT treatment hybridizations is plotted versus the $\log_{10}$ of the expression ratio in the *his3* mutant hybridizations. ORFs that were induced or repressed in both experiments are shown as green and red dots, respectively. The correlation of expression ratios applies not only to genes with large expression ratios (for example, *CHA1* and *ARG1*), but also extends to genes with expression ratios less than 2 (for example, *ILV1* and *CPH1*). *ILV1* is induced 1.9-fold and 1.5-fold, and *CPH1* is downregulated 1.9-fold and 1.7-fold, in cells treated with 3-AT and *his3* mutant cells, respectively. Two ORFs do not fall on the line x = y. The leftmost point is the *HIS3* data point, which is induced by 3-AT treatment but which is not absent from the *his3* mutant strain. The other point is *YOR203w*. Both data points are labeled *HIS3* because hybridization to *YOR203w* is most likely due to *HIS3* mRNA, as *YOR203w* overlaps the *HIS3* open reading frame. ***c***, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from wild-type strain R491 and Cy3-labeled cDNA (green) from strain R1226, deleted for the *HIS3* gene. Arrowheads indicate specific ORFs induced or repressed.

lated to the profile elicited by mutation of the calcineurin genes (ρ = 0.71 ± 0.04), but did not correlate with the expression profile resulting from treatment of the calcineurin mutant strain with CsA (ρ = –0.05 ± 0.07; Table 2), indicating that the genetic deletion of calcineurin interfered with the ability of CsA to elicit its normal transcriptional response. Likewise, the CsA signature was essentially absent in CsA-treated *cph1* mutant cells, and the expression profile of CsA-treated *cph1* mutant cells correlated poorly to that of CsA-treated wild-type cells (ρ = 0.18 ± 0.07). Thus, the *CPH1* gene product was required for the CsA response seen in wild-type cells. Conversely, treatment of *fpr1* mutant cells with CsA resulted in an expression pattern very similar to the profile of CsA-treated wild-type cells (ρ = 0.77 ± 0.03), indicating that *FPR1* was not necessary for the CsA-mediated effects. Analysis of individual ORFs affected by CsA and their expression ratios over the entire set of experiments confirmed that *CPH1* and the genes encoding calcineurin, but not *FPR1*, are necessary for the wild-type CsA response (Fig. 5*b*). The observation that the profiles resulting from FK506 or CsA drug treatment are similar to that of the calcineurin deletion mutant strain might allow the prediction that calcineurin was involved in the pathway affected by these drugs. But because the expression profile of the *fpr1* mutant strain did not bear a strong similarity to the wild-type drug expression profile for FK506, it is obvious that the drug treatment of the mutant strains was necessary to identify Fpr1, but not Cph1, as a potential FK506 drug target. In the same way, the 'decoder' strategy was necessary to identify Cph1, but not Fpr1, as a potential drug target for CsA.

'Decoder' approach can identify secondary drug effects

For a drug that has a single biochemical target, the strategy outlined above may be useful in target validation. In many cases, however, a compound may affect multiple pathways and elicit a very complex signature. 'Decoding' such a complex signature

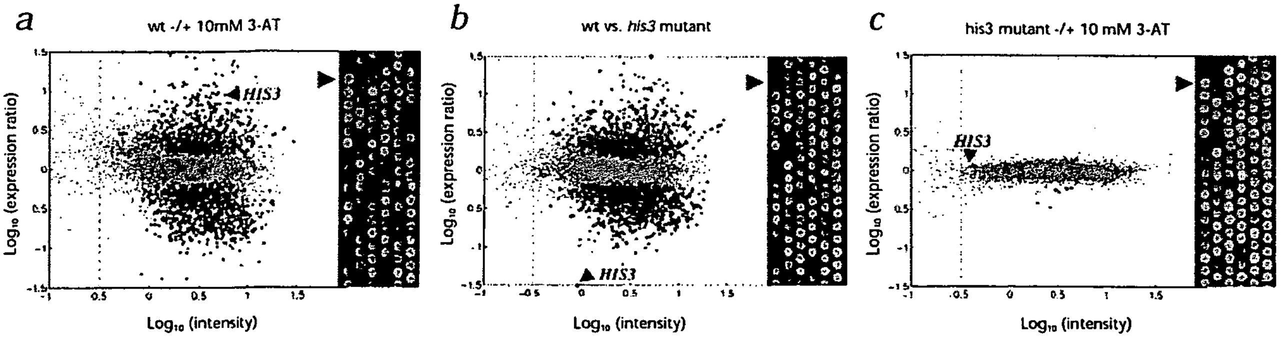


Fig. 4 Treatment of the *his3* mutant strain with 3-AT shows nearly complete loss of 3-AT signature. A plot of the $\log_{10}$ of the mean intensity of hybridization for each ORF versus the $\log_{10}$ of its expression ratio for each experiment is shown next to a pseudo-color image of a representative portion of the microarray. ORFs that are induced or repressed at the 95% confidence level are shown in green and red, respectively. ***a***, Expression profile from treatment of the wild-type (wt) strain with 3-AT. Cy5-labeled cDNA (red) from mock-treated strain R491 and Cy3-labeled cDNA (green) from strain R491 treated with 10 mM 3-AT. ***b***, Expression profile from the *his3* deletion strain. Cy5-labeled cDNA (red) from strain R491 and Cy3-labeled cDNA (green) from strain R1226, deleted for the *HIS3* gene. ***c***, Expression profile of treatment of the *his3* deletion strain with 3-AT. Cy3-labeled cDNA (red) from *his3*-deleted strain R1226 and Cy5-labeled cDNA (green) from strain R1226 treated with 10 mM 3-AT. Arrowheads indicate the DNA probe and data point corresponding to the *HIS3* gene. The blue dashed line represents the threshold below which errors tend to increase rapidly because spot intensities are not sufficiently above background intensity.

Table 2 Signature correlation of expression ratios as a result of CsA treatment in various mutant strains

	wild-type +/–CsA	*cna* +/–CsA	*fpr1* +/–CsA	*cna cph1* +/–CsA	*cph1* +/–CsA
wild-type +/- CsA	0.94 ± 0.04	–0.05 ± .07	0.77 ± 0.03	–0.11 ± 0.07	0.18 ± 0.07

Signature correlation shows the absence of the CsA signature specifically in the calcineurin (*cna*) and *cph1* (cyclophilin) deletion mutants. *cna* represents the mutant with deletions of the catalytic subunits of calcineurin, *CNA1* and *CNA2*. The correlation coefficient reported in the first column represents the correlation between two pairs of hybridizations from independent wild-type +/– CsA experiments.

into the effects mediated through the intended target (the 'on-target signature') and those mediated through unintended targets (the 'off-target' signature) might be useful in evaluating a compound's specificity. Our 'decoder' strategy is based on the premise that 'off-target' signature should be insensitive to the genetic disruption of the primary target.

To determine whether the 'decoder' approach could identify an 'off-target' profile, we looked for a drug-responsive gene whose expression is insensitive to deletion of the primary target. To increase the likelihood of observing such genes, the same strains described in Tables 1 and 2 were treated with higher concentrations (50 µg/ml) of FK506. This led to a much more complex expression profile in wild-type cells, indicating that at this higher concentration, FK506 was inhibiting or activating additional targets. Several of the ORFs in this expanded FK506-induced expression profile were not affected by the calcineurin, *cph1* or *fpr1* mutations, as drug treatment of these mutant strains did not block their presence in the FK506 expression signature (Fig. 6). This indicates that FK506 was triggering changes in transcript levels of many genes through pathways independent of calcineurin, *CPH1* and *FPR1*. Many of the upregulated ORFs in the 'off-target' pathway were genes reported to be regulated by the transcriptional activator Gcn4 (ref. 24). In some strains, a reporter gene under *GCN4* control was induced in response to FK506 treatment[25]. To determine whether *GCN4* is involved in this pathway that is independent of calcineurin, *CPH1* and *FPR1*, we analyzed the effects of treatment with high-dose FK506 on global gene expression in a strain with a *GCN4* deletion (Fig. 6). Of the 41 ORFs with calcineurin-independent expression ratios greater than 4, 32 were not induced in the *gcn4* mutant, indicating that their induction by FK506 was *GCN4*-dependent. Not all *GCN4*-regulated genes were induced by FK506. This FK506-induced subset of *GCN4*-regulated genes may be those most sensitive to subtle changes in Gcn4 levels, or perhaps other regulatory circuits prevent FK506 activation of some *GCN4*-regulated genes. Seven of the remaining nine ORFs induced by FK506 were independent of both the calcineurin and *GCN4* pathways. The simplest explanation is that FK506 inhibits or activates additional pathways. Members of this class include *SNQ2* and *PDR5*, genes that encode drug efflux pumps with structural homology to mammalian multiple drug resistance proteins[26]. FK506 may interact directly with Pdr5 to inhibit its function[27]. Our results indicate that treatment with FK506 leads to fourfold-to-sixfold induction of *PDR5* mRNA levels. *YOR1*, another gene that can confer drug resistance, is also induced threefold-to-fourfold by FK506. Thus, drug treatment of strains with mutations in the primary targets can prove useful in identifying effects mediated by secondary drug targets, including the nature and extent of newly discovered and previously unsuspected pathways affected by the drug.

We describe here a method for drug target validation and the identification of secondary drug target effects that uses DNA microarrays to survey the effects of drugs on global gene expression patterns. We established that genetic and pharmacologic inhibition of gene function can result in extremely similar changes in gene expression. We also demonstrated that one can confirm a potential drug target by treating a deletion mutant defective in the gene encoding the putative target. Drug-mediated signatures from strains with mutations in pathways or processes directly or indirectly affected by the drug bore little or

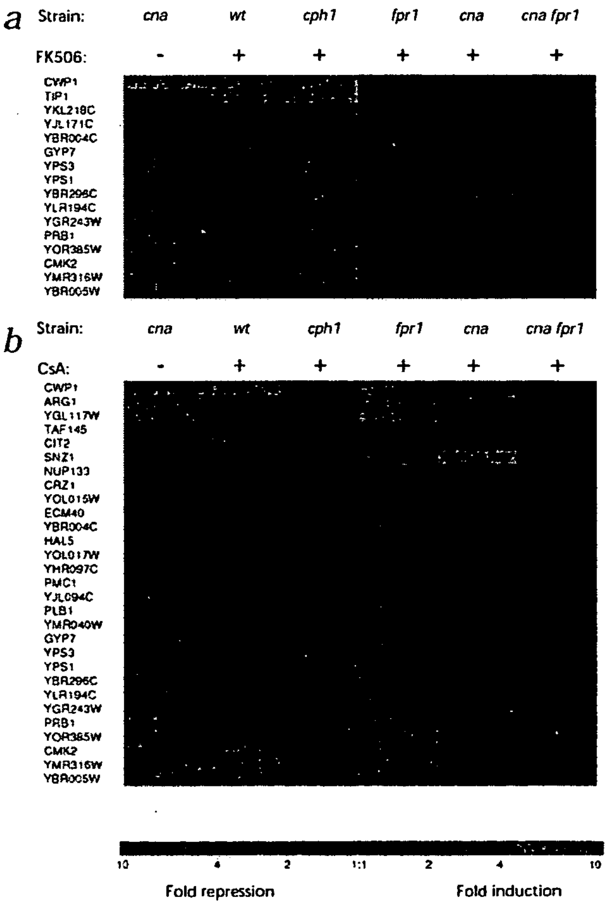


Fig. 5 Response of FK506 and CsA signature genes in strains with deletions in different genes. Genes with expression ratios greater than a factor of 1.8 in response to treatment with 1 µg/ml FK506 (*a*) or 50 µg/ml CsA (*b*) are listed (left side) and their expression ratios in the indicated strain are shown on the green (induction)-red (repression) color scale. ***a***, Calcineurin (*cna*) mutant and FK506 treatment signature genes are in the first two columns. Almost all FK506 signature genes have expression ratios near unity in deletion strains involved in pathways affected by FK506 (calcineurin, *fpr1* and *cna fpr1* mutants) but not in deletion strains in unrelated pathways (*cph1*). ***b***, Calcineurin (*cna*) mutant and CsA treatment signature genes are in the first two columns. Almost all CsA signature genes have expression ratios near unity in deletion strains involved in pathways affected by CsA (calcineurin, *cph1* and *cna cph1* mutants) but not in deletion strains in unrelated pathways (*fpr1*).

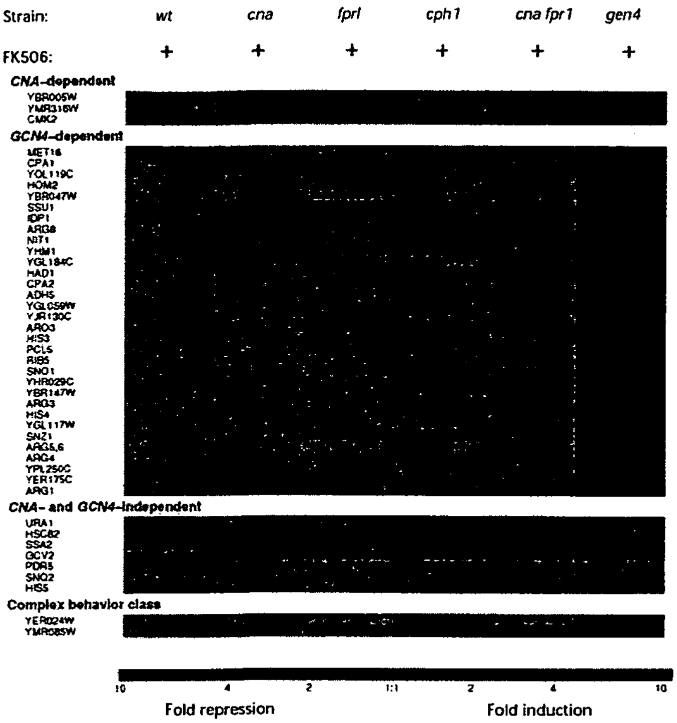


Fig. 6 Response of FK506 signature genes in strains with deletions in different genes. Genes with expression ratios greater than a factor of 4 in at least one experiment are listed and their expression ratios in the indicated strain are shown in the green (induction)–red (repression) color scale. The genes have been divided into classes corresponding to these expected behaviors: '*CNA*-dependent' genes respond to FK506 (50 μg/ml) except when either calcineurin genes or *FPR1* or both are deleted; '*GCN4*-dependent' genes respond to FK506 except when *GCN4* is deleted. These genes still respond to FK506 when calcineurin genes or *FPR1* or *CPH1* are deleted; that is, their responses are not mediated by calcineurin, Cph1, or Fpr1. '*CNA*- and *GCN4*-independent' genes respond to FK506 in all deletion strains tested. A 'complex behavior' class is provided for those genes that did not match the model of FK506 response mediated through calcineurin or Fpr1 or separately through Gcn4.

no similarity to the wild-type drug expression profile. In contrast, drug-mediated signatures from strains with mutations in genes involved in pathways unrelated to the drug's action showed extensive similarity to the wild-type drug signature. By applying this approach to a drug that affects multiple pathways (FK506), we were able to decode a complex signature into component parts, including the identification of an 'off-target' signature that was mediated through pathways independent of calcineurin or the Fpr1 immunophilin.

Discussion

It is well-established that high-throughput biochemical screening can identify potent inhibitory compounds against a given target. The 'decoder' approach described here complements this process by evaluating the equally important property of specificity: the tendency of a compound to inhibit pathways other than that of its intended target. The ability to observe such 'off-target' effects will likely be useful in several ways. Profiling compounds with known toxicities will allow the development of a database of expression changes associated with particular toxicities. Recognition of potential toxicities in the 'off-target' signatures of otherwise promising compounds then may allow earlier identification of those likely to fail in clinical trials. Comparing the extent and peculiarities of 'off-target' signatures of promising drug candiates could provide a new way to group compounds by their effects on secondary pathways, even before those effects are understood. This may prove to be an alternative, potentially more effective, way to select compounds for animal and clinical trials. Some drugs are more effective against a related protein than against the originally intended target. Sildenafil (Viagra™), for example, was initially developed as a phosphodiesterase inhibitor to control cardiac contractility, but was found to be highly specific for phosphodiesterase 5, an isozyme whose inhibition overcomes defects in penile erection. It is possible that application of the 'decoder' to other compounds may show that they too have a potent activity against a target distinct from their intended target.

The ability to decode drug effects is dependent on the availability of functionally 'targetless' cells. In yeast, this is being achieved by systematically disrupting each yeast gene (*Saccharomyces* Deletion Consortium; http://sequence-www.stanford.edu/group/yeast_deletion_project/deletion.html). Efforts are underway to obtain expression profiles from each deletion mutant strain. Determining signatures resulting from inactivation of essential genes presents a unique problem, but it may be possible to do so by examining heterozygotes or by using a controllable promoter to reduce expression of the essential gene. Although it is already feasible to test several compounds in dozens of yeast strains, another challenge for the 'decoder' strategy will be the efficient selection of the mutants with deletions in genes most likely to encode the intended drug target. The signature correlation plots described are one metric that could be used as part of that selection process, but others need to be explored. Applying the 'decoder' to mammalian cells presents additional challenges. It is considerably more difficult to isolate functionally 'targetless' cells. Strategies involving titratable promoters, known specific inhibitors, anti-sense RNAs, ribozymes, and methods of targeting specific proteins for degradation are possible and should be tested. Another limitation is that not all cell types express the same set of genes and therefore 'off-target' effects may be different in different cell types. In addition, applying the 'decoder' to human cells will also require technical improvements that allow expression profiling from a small number of cells. Even the broader question of whether the insensitivity of 'off-target' signatures to the disruption of the main target is the exception or the rule can only be answered by the accumulation of more data. Barkai and Leibler, however, have argued in favor of robustness of biological networks, indicating that drug perturbations ('off-target' signatures) may be robust even when the system is subjected to another perturbation (such as a genetic disruption)(ref. 28). Many practical developments will be necessary if the 'decoder' concept is to be broadly applied.

Expression arrays have been used mainly as an initial screen for genes induced in a particular tissue or process of interest by focusing on genes with large expression ratios. We have found, however, that effort to refine experimental protocols and repeat experiments increases the reliability of the data and permits new applications. For example, it provides a larger set

Table 3 Yeast strains used

Strain	Relevant genotype	Reference
YPH499	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1*	(34)
R563	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 his3::HIS3*	(this study)
R558	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 fpr1::HIS3*	(this study)
R567	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cph1::HIS3*	(this study)
MCY300	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3*	(21)
R132	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3 cph1::kan*^{r}	(this study)
R133	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3 fpr1::kan*^{r}	(this study)
R559	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 his3::HIS3 gcn4::LEU2*	(this study)
BY4719	*Mata trp1-Δ63 ura3-Δ0*	(35)
BY4738	*Matα trp1-Δ 63 ura3-Δ0*	(35)
R491	*Mata/α* BY4719 *X* BY4738	(this study)
BY4728	*Mata his3-Δ200 trp1-Δ63 ura3-Δ0*	(35)
BY4729	*Matα his3-Δ 200 trp1-Δ63 ura3-Δ0*	(35)
R1226	*Mata/α* BY4728 *X* BY4729	(this study)

of genes at higher confidence levels that serve as a more unique signature for a given protein perturbation. In addition, it allows subtle signatures to be detected, when, for example, a protein is only partially inhibited. This may enable clinical monitoring of small changes in protein function in disease or toxicity states before they could otherwise be detected. Because the functions of many genes detected on transcript arrays are known, these microarrays are powerful tools that provide detailed information about a cell's physiology. For example, changes in the flux through a metabolic pathway are reflected in transcriptional changes in genes in the pathway[7]. Furthermore, it may be possible to indirectly measure protein activity levels from expression profiling data (S.F., *et al.*, unpublished data). Thus, although the eventual development of genomic methods allowing the direct measurement of all cellular protein levels will be an important achievement, transcript array technology offers an immediate and robust means of evaluating the effects of various treatments on gene expression and protein function.

Methods

Construction, growth and drug treatment of yeast strains. The strains used in this study (Table 3) were constructed by standard techniques[29]. To construct strain R559, strain R563 was transformed to Leu$^+$ with plasmid pM12 digested by *Sal*I and *Mlu*I (provided by A. Hinnebusch and T. Dever). Strains R132 and R133 were constructed by transforming the bacterial kanamycin resistance cassette[30] flanked by genomic DNA from the *CPH1* and *FPR1* loci, respectively, and selecting for G418-resistant colonies. For experiments with FK506, cells were grown for three generations to a density of 1×10^7 cells/ml in YAPD medium (YPD plus 0.004% adenine) supplemented with 10 mM calcium chloride as described[31]. Where indicated, FK506 was added to a final concentration of 1 µg/ml 0.5 h after inoculation of the culture or to 50 µg/ml 1 h before cells were collected. CsA was used at a final concentration of 50 µg/ml. Cells were broken by standard procedures[32] with the following modifications: Cell pellets were resuspended in breaking buffer (0.2 M Tris HCl pH 7.6, 0.5 M NaCl, 10 mM EDTA, 1% SDS), vortexed for 2 min on a VWR multi-tube vortexer at setting 8 in the presence of 60% glass beads (425–600 µm mesh; Sigma) and phenol:chloroform (50:50, volume/volume). After separation of the phases, the aqueous phase was re-extracted and ethanol-precipitated. Poly A$^+$ RNA was isolated by two sequential chromatographic purifications over oligo dT cellulose (New England Biolabs, Beverly, Massachusetts) using established protocols[32].

For experiments using 3-AT, wild-type or *his3/his3* cells were grown to early logarithmic phase in SC medium, pelleted and resuspended in SC medium lacking histidine for 1 hr in the presence or absence of 10 mM 3-AT, as indicated. Cells were harvested and mRNA isolated as above. FK506 was obtained from the Swedish Hospital Pharmacy (Seattle, Washington) and purified to homogeneity by ethyl acetate extraction by J. Simon (Fred Hutchinson Cancer Research Center, Seattle, Washington). CsA was obtained from Alexis Biochemicals (San Diego, California); 3-AT was from Sigma.

Preparation and hybridization of the labeled sample. Fluorescently-labeled cDNA was prepared, purified and hybridized essentially as described[7]. Cy3- or Cy5-dUTP (Amersham) was incorporated into cDNA during reverse transcription (Superscript II; Life Technologies) and purified by concentrating to less than 10 µl using Microcon-30 microconcentrators (Amicon, Houston, Texas). Paired cDNAs were resuspended in 20–26 µl hybridization solution (3 × SSC, 0.75 µg/ml polyA DNA, 0.2% SDS) and applied to the microarray under a 22- × 30-mm coverslip for 6 h at 63 °C, all according to a published method[7].

Fabrication and scanning of microarrays. PCR products containing common 5′ and 3′ sequences (Research Genetics, Huntsville, Alabama) were used as templates with amino-modified forward primer and unmodified reverse primers to PCR amplify 6,065 ORFs from the *S. cerevisiae* genome. Our first-pass success rate was 94%. Amplification reactions that gave products of unexpected sizes were excluded from subsequent analysis. ORFs that could not be amplified from purchased templates were amplified from genomic DNA. DNA samples from 100-µl reactions were isopropanol-precipitated, resuspended in water, brought to a final concentration of 3x SSC in a total volume of 15 µl, and transferred to 384-well microtiter plates (Genetix Limited, Christchurch, Dorset, England). PCR products were spotted onto 1 × 3-inch polylysine-treated glass slides by a robot built essentially according to defined specifications[3,5,7] (http://cmgm.stanford.edu/pbrown/MGuide). After being printed, slides were processed according to published protocols[7].

Microarrays were imaged on a prototype multi-frame CCD camera in development at Applied Precision (Issaquah, Washington). Each CCD image frame was approximately 2-mm square. Exposure times of 2 s in the Cy5 channel (white light through Chroma 618–648 nm excitation filter, Chroma 657–727 nm emission filter) and 1 s in the Cy3 channel (Chroma 535–560 nm excitation filter, Chroma 570–620 nm emission filter) were done consecutively in each frame before moving to the next, spatially contiguous frame. Color isolation between the Cy3 and Cy5 channels was about 100:1 or better. Frames were 'knitted' together in software to make the complete images. The intensity of spots (about 100 µm) were quantified from the 10-µm pixels by frame-by-frame background subtraction and intensity averaging in each channel. Dynamic range of the resulting spot intensities was typically a ratio of 1,000 between the brightest spots and the background-subtracted additive error level. Normalization between the channels was accomplished by normalizing each channel to the mean intensities of all genes. This procedure is nearly equivalent to normalization between channels using the intensity ratio of genomic DNA spots[7], but is possibly more robust, as it is based on the intensities of several thousand spots distributed over the array.

Signature correlation coefficients and their confidence limits. Correlation coefficients between the signature ORFs of various experiments were calculated using:

$$\rho = \sum_k x_k y_k \,/\, (\sum_k x_k^2 \sum_k y_k^2)^{1/2}$$

where x_k is the $\log_{10}$ of the expression ratio for the k^{th} gene in the x signature, and y_k is the $\log_{10}$ of the expression ratio for the k^{th} gene in the y signature. The summation is over those genes that were either up- or down-regulated in either experiment at the 95% confidence level. These genes each had a less than 5% chance of being actually unregulated (having expression ratios departing from unity due to measurement errors alone). This confidence level was assigned based on an error model which assigns a lognormal probability distribution to each gene's expression ratio with characteristic width based on the observed scatter in its repeated measurements (repeated arrays at the same nominal experimental conditions) and on the individual array hybridization quality. This latter dependence was derived from control experiments in which both Cy3 and Cy5 samples were derived from the same RNA sample. For large numbers of repeated measurements the error reduces to the observed scatter. For a single measurement the error is based on the array quality and the spot intensity.

Random measurement errors in the x and y signatures tend to bias the correlation towards zero. In most experiments, most genes are not significantly affected but do show small random measurement errors. Selecting only the '95% confidence' genes for the correlation calculation, rather than the entire genome, reduces this bias and makes the actual biological correlations more apparent.

Correlations between a profile and itself are unity by definition. Error limits on the correlation are 95% confidence limits based on the individual measurement error bars, and assuming uncorrelated errors[33]. They do not include the bias mentioned above; thus, a departure of ρ from unity does not necessarily mean that the underlying biological correlation is imperfect. However, a correlation of 0.7 ± 0.1, for example, is very significantly different from zero. Small (magnitude of $\rho < 0.2$) but formally significant correlation in the tables and text probably are due to small systematic biases in the Cy5/Cy3 ratios that violate the assumption of independent measurement errors used to generate the 95% confidence limits. Therefore, these small correlation values should be treated as not significant. A likely source of uncorrected systematic bias is the partially corrected scanner detector nonlinearity that differently affects the Cy3 and Cy5 detection channels.

The 1 µg/ml FK506 treatment signature was compared with more than 40 unrelated deletion mutant strain or drug signatures. These control profiles had correlation coefficients with the FK506 profile that were distributed around zero (mean $\rho = -0.03$) with a standard deviation of 0.16 (data not shown), and none had correlations greater than $\rho = 0.38$. Similarly, the calcineurin mutant strain signature correlated well with the CsA treatment signature ($\rho = 0.71 \pm 0.04$) but not with the signatures from the negative controls (mean $\rho = -0.02$ with a standard deviation of 0.18).

Quality controls. End-to-end checks on expression ratio measurement accuracy were provided by analyzing the variance in repeated hybridizations using the same mRNA labeled with both Cy3 and Cy5, and also using Cy3 and Cy5 mRNA samples isolated from independent cultures of the same nominal strain and conditions. Biases undetected with this procedure, such as gene-specific biases presumably due to differential incorporation of Cy3- and Cy5-dUTP into cDNA, were minimized by doing hybridizations in fluor-reversed pairs, in which the Cy3/Cy5 labeling of the biological conditions was reversed in one experiment with respect to the other. The expression ratio for each gene is then the ratio of ratios between the two experiments in the pair. Other biases are removed by algorithmic numerical de-trending. The magnitude of these biases in the absence of de-trending and fluor reversal is typically about 30% in the ratio, but may be as high as twofold for some ORFs.

Expression ratios are based on mean intensities over each spot. Some smaller spots have fewer image pixels in the average. This does not degrade accuracy noticeably until the number of pixels falls below ten, in which case the spot is rejected from the data set. 'Wander' of spot positions with respect to the nominal grid is adaptively tracked in array subregions by the image processing software. Unequal spot 'wander' within a subregion greater than half-a-spot spacing is a difficulty for the automated quantitating algorithms; in this case, the spot is rejected from analysis based on human inspection of the 'wander'. Any spots partially overlapping are excluded from the data set. Less than 1% of spots typically are rejected for these reasons.

Acknowledgments

The authors thank all the members of Rosetta for their contributions to this work. We thank P. Linsley, D. Shoemaker and A. Murray for critical reading of the manuscript, and M. Cyert for providing yeast strains. Work done at Stanford was supported in part by the Howard Hughes Medical Institute, and by a grant to P.O.B from the NHGRI. P.O.B is an assistant investigator of the Howard Hughes Medical Institute.

RECEIVED 13 AUGUST; ACCEPTED 2 OCTOBER 1998

1. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
2. Schena, M. *et al.* Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* **93**, 10614–10619 (1996).
3. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645 (1996).
4. Lockhart, D.J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).
5. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* **14**, 457–460 (1996).
6. Heller, R.A. *et al.* Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA* **94**, 2150–2155 (1997).
7. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
8. Lashkari, D.A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* **94**, 13057–13062 (1997).
9. Wodicka, L., Dong H., Mittman, M, Ho, M.-H. & Lockhart, D.J.. Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nature Biotechnol.* **15**, 1359–1367 (1997).
10. Cho, R.J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73 (1998).
11. Gray, N.S. *et al.* Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* **281**, 533–538 (1998).
12. Cardenas, M.E., Lorenz, M., Hemenway, C. & Heitman, J. Yeast as model T cells. *Perspect. Drug Discovery Design* **2**, 103–126 (1994).
13. Klee, C.B., Ren, H. & Wang, X. Regulation of the calmodulin-stimulated protein phosphatase, calcineurin. *J. Biol. Chem.* **273**, 13367–13370 (1998).
14. Tanida, I., Hasegawa, A., Iida, H., Ohya, Y. & Anraku, Y. Cooperation of calcineurin and vacuolar H(+)-ATPase in intracellular Ca2+ homeostasis of yeast cells. *J. Biol. Chem.* **270**, 10113–10119 (1995).
15. Moser, M.J., Geiser, J.R. & Davis, T.N. Ca2+-calmodulin promotes survival of pheromone-induced growth arrest by activation of calcineurin and Ca2+-calmodulin-dependent protein kinase. *Mol. Cell. Biol.* **16**, 4824–4831 (1996).
16. Mizunuma, M., Hirata, D., Miyahara, K., Tsuchiya, E. & Miyakawa, T. Role of calcineurin and Mpk1 in regulating the onset of mitosis in budding yeast. *Nature* **392**, 303–306 (1998).
17. Yazdanbakhsh, K., Choi, J.W., Li, Y., Lau, L.F. & Choi, Y. Cyclosporin A blocks apoptosis by inhibiting the DNA binding activity of the transcription factor Nur77. *Proc. Natl. Acad. Sci. USA* **92**, 437–441 (1995).
18. Molkentin, J.D. *et al.* A calcineurin-dependent transcriptional pathway for cardiac hypertrophy. *Cell* **93**, 215–228 (1998)
19. Mansuy, I.M., Mayford, M., Jacob, B., Kandel, E.R. & Bach, M.E. Restricted and regulated overexpression reveals calcineurin as a key component in the transition from short-term to long-term memory. *Cell* **92**, 39–49 (1998).
20. Schreiber, S.L. & Crabtree, G.R. The mechanism of action of cyclosporin A and FK506. *Immunol. Today* **13**, 136–142 (1992).
21. Cyert, M.S., Kunisawa, R., Kaim, D. & Thorner, J. Yeast has homologs (CNA1 and CNA2 gene products) of mammalian calcineurin, a calmodulin-regulated phosphoprotein phosphatase. *Proc. Natl. Acad. Sci. USA* **88**, 7376–7380 (1991).
22. Jones, E.W. & Fink, G.R. in *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression* (eds. Strathern, J.N., Jones, E.W. & Broach, J.R.) 181–299 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1982).
23. Hinnebusch, A. Translational regulation of yeast GCN4. *J. Biol. Chem.* **272**, 21661–21664 (1997).
24. Hinnebusch, A.G. in *The Molecular and Cellular Biology of the Yeast*

Saccharomyces: Gene Expression. (eds. Jones, E.W., Pringle, J.R. & Broach, J.R.) 319-414 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1992).

25. Heitman, J. *et al.* The immunosuppressant FK506 inhibits amino acid import in Saccharomyces cerevisiae. *Mol. Cell. Biol.* 13, 5010-5019 (1993).
26. Balzi, E. & Goffeau, A. Yeast multidrug resistance: the PDR network. *J. Bioenerg. Biomembr.* 27, 71-76 (1995).
27. Egner, R., Rosenthal, F.E., Kralli, A., Sanglard, D. & Kuchler, K. Genetic separation of FK506 susceptibility and drug transport in the yeast Pdr5 ATP-binding cassette multidrug resistance transporter *Mol. Biol. Cell* 9, 523-543 (1998).
28. Barkal, N. & Leiber, S. Robustness in simple biochemical networks. *Nature* 387, 913-917 (1997).
29. Schiestl, R.H., Manivasakam, P, Woods, R.A. & Gietz, R.D. Introducing DNA into yeast by transformation. *Methods: A companion to Methods in Enzymology* 5, 79-85 (1993).
30. Wach, A., Brachat, A., Pohlmann, R. & Philippsen, P. New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. *Yeast* 10, 1793-1808 (1994).
31. Garrett-Engele, P., Moilanen, B. & Cyert, M.S . Calcineurin, the Ca2+/calmodulin-dependent protein phosphatase, is essential in yeast mutants with cell integrity defects and in mutants that lack a functional vacuolar H(+)-ATPase. *Mol. Cell. Biol.* 15, 4103-4114 (1995).
32. Ausubel, F.M. *et al.* in *Current Protocols in Molecular Biology* 13.12.1-13.12.5 (eds. Ausubel, F.M., *et al.*)(John Wiley & Sons, New York, 1993).
33. Bulmer, M.G. in *Principles of Statistics* 224-225 (Dover Publications, New York, 1979).
34. Sikorski, R.S. & Hieter, P. A system of shuttle vectors and yeast host strains designated for efficient manipulation of DNA in *Saccharomyces cerevisiae. Genetics* 122, 19-27 (1989).
35. Brachmann, C.B. *et al.* Designer deletion strains derived from Saccharomyces cerevisiae S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* 14, 115-132 (1998).

Exhibit D of Iyer Declaration
with Response dated 4/23/04
In USSN: 10/049,742

co mosaic viral RNA was obtained by phenol and chloroform extractions of the virus and precipitated from ethanol. CA-NC assembly reactions in the presence of noncognate RNAs were identical to those given in (9). In the absence of RNA, CA-NC cones formed under the following conditions: 300 μM CA-NC, 1 M NaCl, and 50 mM tris-HCl (pH 8.0) at 37°C for 60 min. In the absence of exogenous RNA, neither cones nor cylinders formed at concentrations of 0.5 M NaCl or below. Absorption spectra demonstrated that our CA-NC preparations were not contaminated with *Escherichia coli* RNA (estimated lower detection limit was $\sim$1 base/protein molecule). To control for even lower levels of RNA contamination, we preincubated the CA-NC protein with 0.5 mg/ml ribonuclease A (Type 1-AS, 54 Kunitz U/mg, Sigma) for 1 hour at 4°C, which then formed cones normally.

13. V. Y. Klishko, data not shown.
14. M. Ge and K. Sattler, *Chem. Phys. Lett.* **220**, 192 (1994).
15. A. Krishnan *et al.*, *Nature* **388**, 451 (1997).
16. L. B. Kong *et al.*, *J. Virol.* **72**, 4403 (1998).
17. Assembly mixtures were deposited on holey carbon grids, blotted briefly with filter paper, plunged into liquid ethane, and transferred to liquid nitrogen. Frozen grids were transferred to a Philips 420 TEM equipped with a Gatan cold stage system, and images of particles in vitreous ice were recorded under low dose conditions at 36,000× magnification and $\sim$1.6-μm defocus.
18. J. T. Finch, data not shown.
19. R. A. Crowther, *Proceedings of the Third John Innes Symposium* (1976), pp. 15–25; E. Kellenberger, M. Häner, M. Wurtz, *Ultramicroscopy* **9**, 139 (1982); J. Seymore and D. J. DeRosier, *J. Microsc.* **148**, 195 (1987).
20. M. V. Nermut, C. Grief, S. Hashmi, D. J. Hockley, *AIDS Res. Hum. Retroviruses* **9**, 929 (1993); M. V. Nermut *et al.*, *Virology* **198**, 288 (1994); E. Barklis, J. McDermott, S. Wilkens, S. Fuller, D. Thompson, *J. Biol. Chem.* **273**, 7177 (1998); E. Barklis *et al.*, *EMBO J.* **16**, 1199 (1997); M. Yeager, E. M. Wilson-Kubalek, S. G. Weiner, P. O. Brown, A. Rein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 7299 (1998).
21. J. T. Finch *et al.*, unpublished observations.
22. V. M. Vogt, in (2), pp. 27–70.
23. M. A. McClure, M. S. Johnson, D.-F. Feng, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2469-2473 (1988).
24. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

25. We thank C. Hill for very helpful discussions on the relationship between viral cores and fullerene cones, D. Hobbs for refining the ChemDraw3D images of cones, G. Stubbs for a gift of tobacco mosiac virus, J. McCutcheon for the plasmid used to prepare ribosomal RNA, and K. Albertine and N. Chandler of the University of Utah Shared Electron Microscopy facility for their support and encouragement. Supported by grants from NIH and from the Huntsman Cancer Institute (to W.I.S.).

29 September 1998; accepted 17 November 1998

The Transcriptional Program in the Response of Human Fibroblasts to Serum

Vishwanath R. Iyer, Michael B. Eisen, Douglas T. Ross, Greg Schuler, Troy Moore, Jeffrey C. F. Lee, Jeffrey M. Trent, Louis M. Staudt, James Hudson Jr., Mark S. Boguski, Deval Lashkari, Dari Shalon, David Botstein, Patrick O. Brown*

The temporal program of gene expression during a model physiological response of human cells, the response of fibroblasts to serum, was explored with a complementary DNA microarray representing about 8600 different human genes. Genes could be clustered into groups on the basis of their temporal patterns of expression in this program. Many features of the transcriptional program appeared to be related to the physiology of wound repair, suggesting that fibroblasts play a larger and richer role in this complex multicellular response than had previously been appreciated.

The response of mammalian fibroblasts to serum has been used as a model for studying growth control and cell cycle progression (*1*). Normal human fibroblasts require growth factors for proliferation in culture; these growth factors are usually provided by fetal bovine serum (FBS). In the absence of growth factors, fibroblasts enter a nondividing state, termed G_0, characterized by low metabolic activity. Addition of FBS or purified growth factors induces proliferation of the fibroblasts; the changes in gene expression that accompany this proliferative response have been the subject of many studies, and the responses of dozens of genes to serum have been characterized.

We took a fresh look at the response of human fibroblasts to serum, using cDNA microarrays representing about 8600 distinct human genes to observe the temporal program of transcription that underlies this response. Primary cultured fibroblasts from human neonatal foreskin were induced to enter a quiescent state by serum deprivation for 48 hours and then stimulated by addition of medium containing 10% FBS (*2*). DNA microarray hybridization was used to measure the temporal changes in mRNA levels of 8613 human genes (*3*) at 12 times, ranging from 15 min to 24 hours after serum stimulation. The cDNA made from purified mRNA from each sample was labeled with the fluorescent dye Cy5 and mixed with a common reference probe consisting of cDNA made from purified mRNA from the quiescent

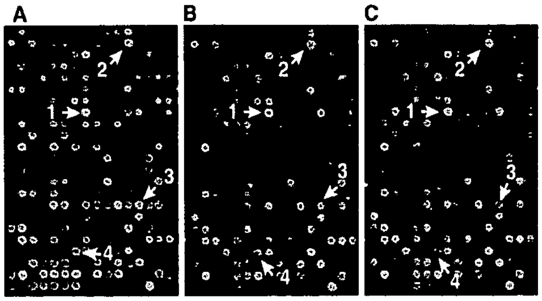


Fig. 1. The same section of the microarray is shown for three independent hybridizations comparing RNA isolated at the 8-hour time point after serum treatment to RNA from serum-deprived cells. Each microarray contained 9996 elements, including 9804 human cDNAs, representing 8613 different genes. mRNA from serum-deprived cells was used to prepare cDNA labeled with Cy3–deoxyuridine triphosphate (dUTP), and mRNA harvested from cells at different times after serum stimulation was used to prepare cDNA labeled with Cy5-dUTP. The two cDNA probes were mixed and simultaneously hybridized to the microarray. The image of the subsequent scan shows genes whose mRNAs are more abundant in the serum-deprived fibroblasts (that is, suppressed by serum treatment) as green spots and genes whose mRNAs are more abundant in the serum-treated fibroblasts as red spots. Yellow spots represent genes whose expression does not vary substantially between the two samples. The arrows indicate the spots representing the following genes: 1, protein disulfide isomerase-related protein P5; 2, IL-8 precursor; 3, EST AA057170; and 4, vascular endothelial growth factor.

V. R. Iyer and D. T. Ross, Department of Biochemistry, Stanford University School of Medicine, Stanford CA 94305, USA. M. B. Eisen and D. Botstein, Department of Genetics, Stanford University School of Medicine, Stanford CA 94305, USA. G. Schuler and M. S. Boguski, National Center for Biotechnology Information, Bethesda MD 20894, USA. T. Moore and J. Hudson Jr., Research Genetics, Huntsville, AL 35801, USA. J. C. F. Lee, D. Lashkari, D. Shalon, Incyte Pharmaceuticals, Fremont, CA 94555, USA. J. M. Trent, Laboratory of Cancer Genetics, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA. L. M. Staudt, Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, Bethesda, MD 20892, USA. P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford CA 94305, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

culture (time zero) labeled with a second fluorescent dye, Cy3 (*4*). The color images of the hybridization results (Fig. 1) were made by representing the Cy3 fluorescent image as green and the Cy5 fluorescent image as red and merging the two color images.

Diverse temporal profiles of gene expression could be seen among the 8613 genes surveyed in this experiment (Fig. 2); many of these genes (about half) were unnamed expressed sequence tags (ESTs) (*5*). Although diverse patterns of expression were observed, the orderly choreography of the expression program became apparent when the results were analyzed by a clustering and display method developed in our laboratory for analyzing genome-wide gene expression data (*6*). An example of such an analysis, here applied to a subset of 517 genes whose expression changed substantially in response to serum (*7*), is shown in Fig. 2. The entire detailed data set underlying Fig. 2 is available as a tab-delimited table (in cluster order) at the *Science* Web site (www.sciencemag.org/feature/data/984559.shl). In addition, the entire, larger data set for the complete set of genes analyzed in this experiment can be found at a Web site maintained by our laboratory (genome-www.stanford.edu/serum) (*8*).

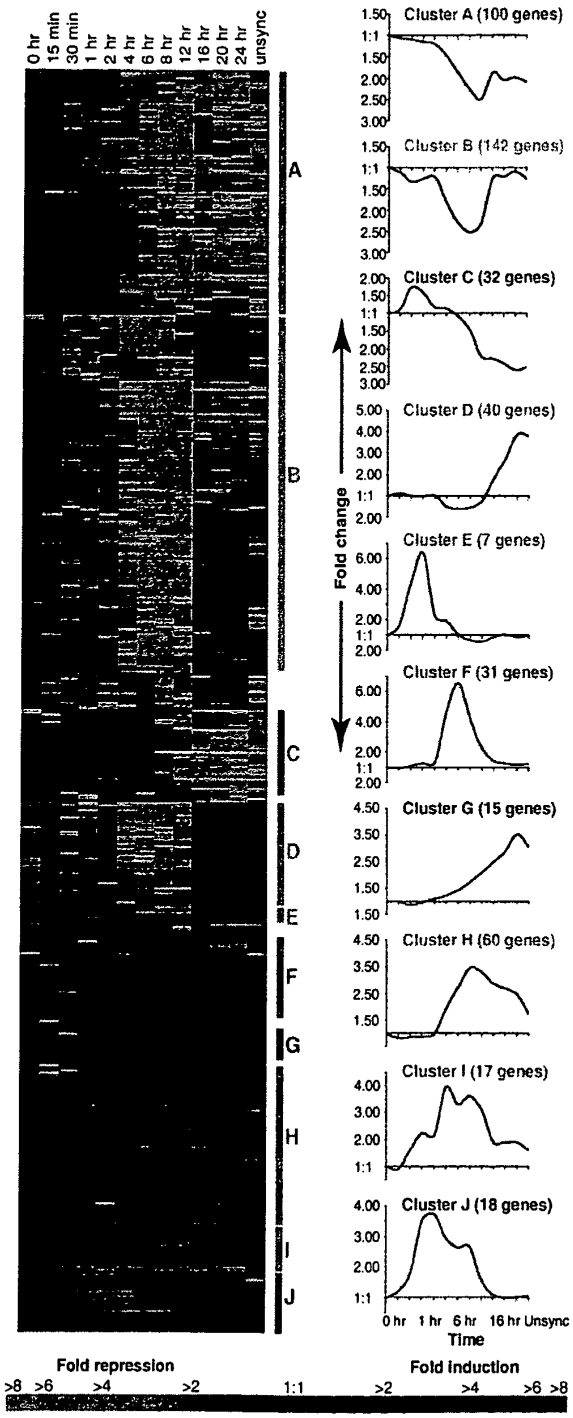


Fig. 2. Cluster image showing the different classes of gene expression profiles. Five hundred seventeen genes whose mRNA levels changed in response to serum stimulation were selected (*7*). This subset of genes was clustered hierarchically into groups on the basis of the similarity of their expression profiles by the procedure of Eisen *et al.* (*6*). The expression pattern of each gene in this set is displayed here as a horizontal strip. For each gene, the ratio of mRNA levels in fibroblasts at the indicated time after serum stimulation ("unsync" denotes exponentially growing cells) to its level in the serum-deprived (time zero) fibroblasts is represented by a color, according to the color scale at the bottom. The graphs show the average expression profiles for the genes in the corresponding "cluster" (indicated by the letters A to J and color coding). In every case examined, when a gene was represented by more than one array element, the multiple representations in this set were seen to have identical or very similar expression profiles, and the profiles corresponding to these independent measurements clustered either adjacent or very close to each other, pointing to the robustness of the clustering algorithm in grouping genes with very similar patterns of expression.

One measure of the reliability of the changes we observed is inherent in the expression profiles of the genes. For most genes whose expression levels changed, we could see a gradual change over a few time points, which thus effectively provided independent measurements for almost all of the observations. An additional check was provided by the inclusion of duplicate and, in a few cases, multiple array elements representing the same gene for about 5% of the genes included in this microarray. In addition, three independent hybridizations to different microarrays with mRNA samples from cells harvested 8 hours after serum addition showed good correlation (Fig. 1). As an independent test, we measured the expression levels of several genes using the TaqMan 5′ nuclease fluorigenic quantitative polymerase chain reaction (PCR) assay (*9*). The expression profiles of the genes, as measured by these two independent methods, were very similar (Fig. 3) (*10*).

The transcriptional response of fibroblasts to serum was extremely rapid. The immediate response to serum stimulation was dominated by genes that encode transcription factors and other proteins involved in signal transduction. The mRNAs for several genes [including c-FOS, JUN B, and mitogen-activated protein (MAP) kinase phosphatase–1 (MKP1)] were detectably induced within 15 min after serum stimulation (Fig. 4, A and B). Fifteen of the genes that were observed to be induced by serum encode known or suspected regulators of transcription (Fig. 4B). All but one were immediate-early genes—their induction was not inhibited by cycloheximide (*11*). This class of genes could be distinguished into those whose induction was transient (Fig. 2, cluster E) and those whose mRNA levels remained induced for much longer (Fig. 2, clusters I and J). Some features of the immediate response appeared to be directed at adaptation to the initiating signals. We observed a marked induction of mRNA encoding MKP1, a dual-specificity phosphatase that modulates the activity of the ERK1 and ERK2 MAP kinases (*12*). The coincidence of the peak of expression of genes in cluster E (Fig. 2) with that of MKP1 (Fig. 4A) suggests the possibility that continued activity of the MAP kinase pathway is required to maintain induction of these genes but not of those with sustained expression (clusters I and J). The gene encoding a second member of the dual-specificity MAP kinase phosphatase family, known as dual-specificity protein phosphatase 6/pyst2, was induced later, at about 4 hours after serum stimulation. Genes encoding diverse other proteins with roles in signal transduction, ranging from cell-surface receptors [for example, the sphingosine 1–phosphate receptor (EDG-1), the vascular endothelial growth factor receptor, and the type II BMP receptor] to regulators of G-protein signaling (for example, NET1/p115 rho GEF) to DNA-binding transcription factors, were induced by serum (Fig. 4A).

The reprogramming of the regulatory circuits in response to serum involved not only induction of transcription factors but also reduced expression of many transcriptional regulators—some of which may play roles in maintaining the cells in G_0 or in priming them to react to wounding (Fig. 4C). Perhaps as a consequence of the historical focus on genes induced by serum stimulation of fibroblasts, the set of transcription factors whose expression diminished upon serum stimulation has been less well characterized.

Genes known or likely to be involved in controlling and mediating the proliferative response showed distinctive patterns of regulation. Several genes whose products inhibit progression of the cell-division cycle, such as p27 Kip1, p57 Kip2, and p18, were expressed in the quiescent fibroblasts and down-regulated before the onset of cell division. The nadir in the mRNA levels for these genes occurred between 6 and 12 hours after serum stimulation (Fig. 5A), coincident with the passage of the fibroblasts through G_1. The levels of the transcript encoding the WEE1-like protein kinase, which is believed to inhibit mitosis by phosphorylation of Cdc2, diminished between 4 and 8 to 12 hours after serum addition (Fig. 5A), well before the onset of M phase at around 16 hours, raising the possibility of an additional role for Wee1 in an earlier stage of the cell cycle or in regulating the G_0 to G_1 transition. Several genes induced in the first few hours after serum stimulation, such as the helix-loop-helix proteins ID2 and ID3 and EST AA016305, a gene with homology to G_1-S cyclins, are candidates for roles in promoting the exit from G_0.

Genes involved in mediating progression through the cell cycle were characterized by a distinctive pattern of expression (Fig. 2, cluster D), reflecting the coincidence of their expression with the reentry of the stimulated fibroblasts into the cell-division cycle. The stimulated fibroblasts replicated their DNA about 16 hours after serum treatment. This timing was reflected by the induction of mRNA encoding both subunits of ribonucleotide reductase and PCNA, the processivity factor for DNA polymerase epsilon and delta. Cyclin A, Cyclin B1, Cdc2, and CDC28 kinase, regulators of passage through the S phase and the transition from G_2 to M phase, were induced at about 16 to 20 hours after serum addition. The kinase in the Cyclin B1–CDK pair needs to be activated by phosphorylation. The gene encoding Cyclin-dependent kinase 7 (CDK7; a homolog of *Xenopus* MO15 cdk-activating kinase) was induced in parallel with the Cdc2 and Cdc28 kinases (Fig. 5A), suggesting a potential role for CDK7 in mediating M phase. DNA topoisomerase II α, required for chromosome segregation at mitosis; Mad2, a component of the spindle checkpoint that prevents completion of mitosis (anaphase) if chromosomes are not attached to the spindle; and the kinetochore protein CENP-F all showed a similar expression profile.

In the hours after the serum stimulus, one of the most striking features of the unfolding transcriptional program was the appearance of numerous genes with known roles in processes relevant to the physiology of wound healing. These included both genes involved in the direct role played by fibroblasts in remodeling of the clot and the extracellular matrix and, more notably, genes encoding proteins involved in intercellular signaling (Fig. 5). Genes induced in this program encode products that can (i) participate in the dynamic process of clotting, clot dissolution, and remodeling and perhaps contribute to hemostasis by promoting local vasoconstriction (for example, endothelin-1); (ii) promote chemotaxis and activation of neutrophils (for example, COX2) and recruitment and extravasation of monocytes and macrophages (for example, MCP1); (iii) promote chemotaxis and activation of T lymphocytes [for example, interleukin-8 (IL-8)] and B lymphocytes (for example, ICAM-1), thus providing both innate and antigen-specific defenses against wound infection and recruiting the phagocytic cells that will be required to clear out the debris during remodeling of the wound; (iv) promote angiogenesis and neovascularization (for example, VEGF) through newly forming tissue; (v) promote migration and proliferation of fibroblasts (for example, CTGF) and their differentiation into myofibroblasts (for example, Vimentin); and (vi) promote migration and proliferation of keratinocytes, leading to reepithelialization of the wound (for example, FGF7), and promote proliferation of melanocytes, perhaps contributing to wound hyperpigmentation (for example, FGF2).

Coordinated regulation of groups of genes whose products act at different steps in a common process was a recurring theme. For example, Furin, a prohormone-processing protease required for one of the processing steps in the generation of active endothelin, was induced in parallel with induction of the gene encoding the precursor of endothelin-1 (Fig. 5E) ([13]). Conversely, expression of CALLA/CD10, a membrane metalloprotease that degrades endothelin-1 and other peptide mediators of acute inflammation, was re-

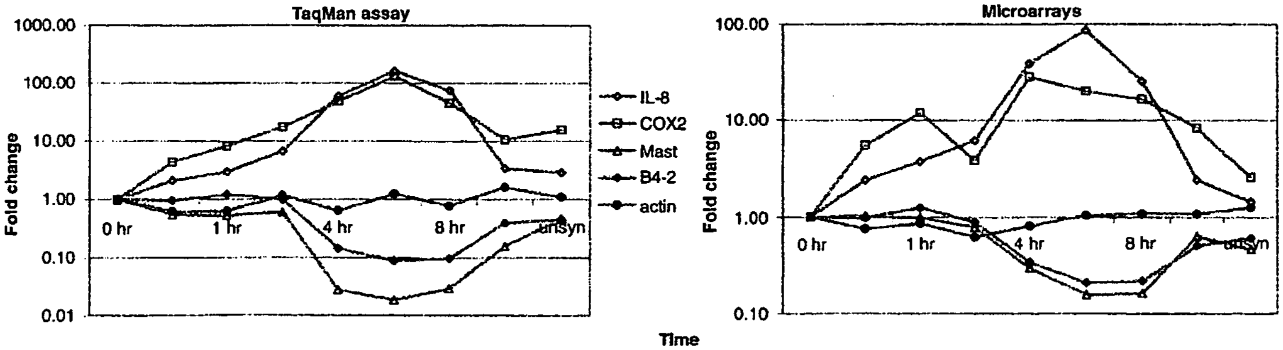


Fig. 3. Independent verification of microarray quantitation. Relative mRNA levels of the indicated genes (Mast, mast/stem cell growth factor receptor) were measured with the TaqMan 5′ nuclease fluorigenic quantitative PCR assay (9) (left) in the same samples that were used to prepare probes for microarray hybridizations (right). Data from the TaqMan analysis were normalized to mRNA concentrations and plotted relative to the level at time zero, so that the results could be compared with those from the microarray hybridizations. In general, quantitation with the two methods gave very similar results ([10]).

duced. A second example is provided by a set of five genes involved in the biosynthesis of cholesterol (Fig. 5I). The mRNAs encoding each of these enzymes showed sharply diminished expression beginning 4 to 6 hours after serum stimulation of fibroblasts. A likely explanation for the coordinated down-regulation of the cholesterol biosynthetic pathway is that serum provides cholesterol to fibroblasts through low-density lipoproteins, whereas in the absence of the cholesterol provided by serum, endogenous cholesterol biosynthesis in fibroblasts is required.

Many of the previously studied genes that we observed to be regulated in this program have no recognized role in any aspect of wound healing or fibroblast proliferation. Their identification in this study may therefore point to previously unknown aspects of these processes. A few selected genes in this group are shown in Fig. 5H. The stanniocalcin gene, for example (Fig. 5H), encodes a secreted protein without a clearly identified function in human cells (*14*, *15*). Its induction in serum-stimulated fibroblasts suggests the possibility that it may play a role in the wound-healing process, perhaps serving as a signal in mediating inflammation or angiogenesis.

One of the most important results of this exploration was the discovery of over 200 previously unknown genes whose expression was regulated in specific temporal patterns during the response of fibroblasts to serum. For example, 13 of the 40 genes in cluster D (Fig. 2) have descriptive names that reflect their putative function. Nine of these 13 genes (69%) encode proteins that play roles in cell cycle progression, particularly in DNA replication and the G_2-M transition. This enrichment for cell cycle–related genes suggests that some of the unnamed genes in this cluster—for example, EST W79311 and EST R13146, neither of which have sequence similarity to previously characterized genes—may represent previously unknown genes involved in this part of the cell cycle. Similarly, a remarkable fraction of genes that were grouped into cluster F on the basis of their expression profiles encoded proteins involved in intercellular signaling (Fig. 2), suggesting that a similar role should be considered for the many unnamed genes in this cluster. A disproportionately large fraction of the genes whose transcription diminished upon serum stimulation were unnamed ESTs.

Our intention was to use this experiment as a model to study the control of the transition

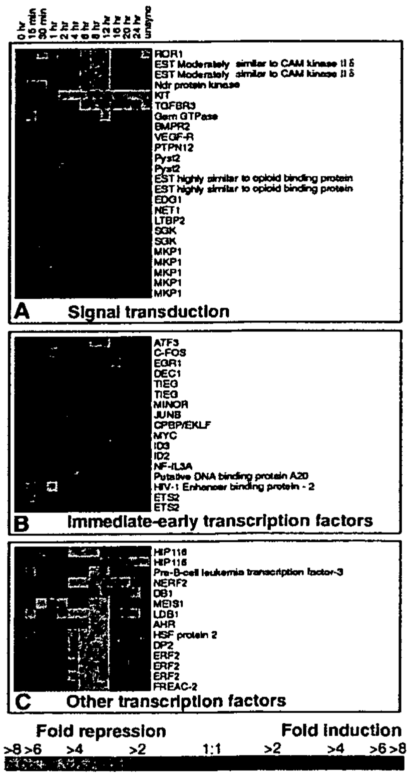


Fig. 4. "Reprogramming" of fibroblasts. Expression profiles of genes whose function is likely to play a role in the reprogramming phase of the response are shown with the same representation as in Fig. 2. In the cases in which a gene was represented by more than one element in the microarray, all measurements are shown. The genes were grouped into categories on the basis of our knowledge of their most likely role. Some genes with pleiotropic roles were included in more than one category.

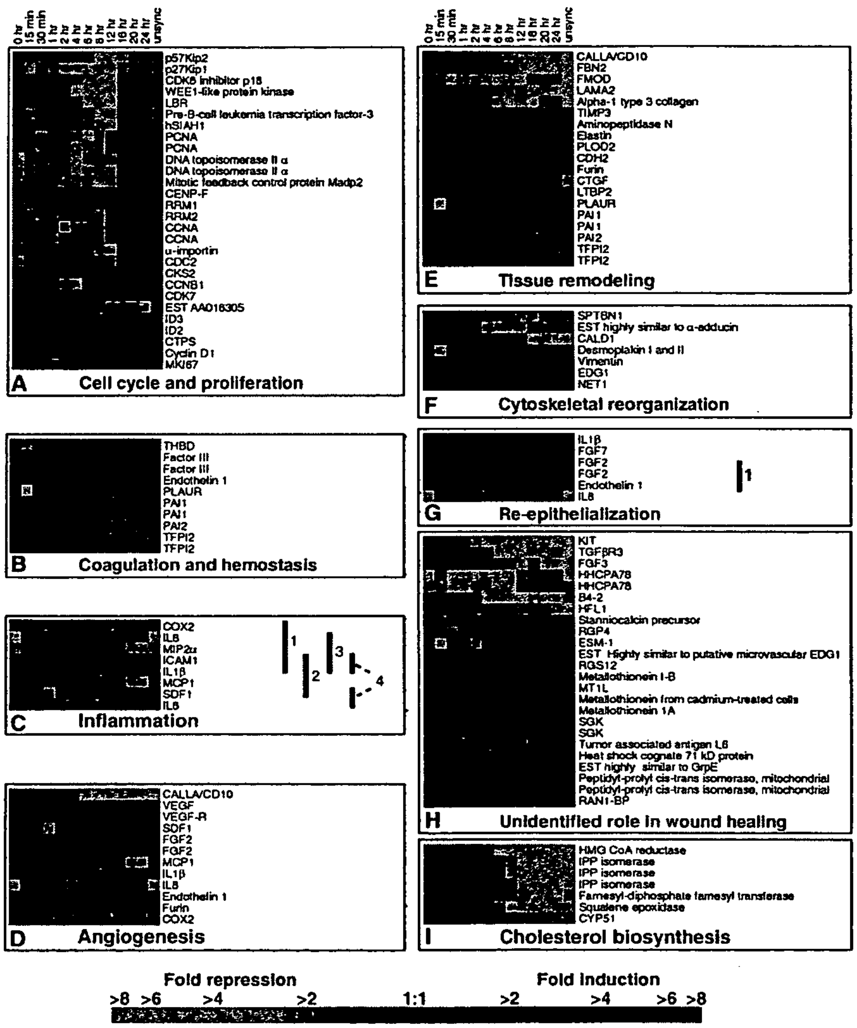


Fig. 5. The transcriptional response to serum suggests a multifaceted role for fibroblasts in the physiology of wound healing. The features of the transcriptional program of fibroblasts in response to serum stimulation that appear to be related to various aspects of the wound-healing process and fibroblast proliferation are shown with the same convention for representing changes in transcript levels as was used in Figs. 2 and 4. (**A**) Cell cycle and proliferation, (**B**) coagulation and hemostasis, (**C**) inflammation, (**D**) angiogenesis, (**E**) tissue remodeling, (**F**) cytoskeletal reorganization, (**G**) reepithelialization, (**H**) unidentified role in wound healing, and (**I**) cholesterol biosynthesis. The numbers in (C) and (G) refer to genes whose products serve as signals to neutrophils (C1), monocytes and macrophages (C2), T lymphocytes (C3), B lymphocytes (C4), and melanocytes (G1).

from G_0 to a proliferating state. However, one of the defining characteristics of genome-scale expression profiling experiments is that the examination of so many diverse genes opens a window on all the processes that actually occur and not merely the single process one intended to observe. Serum, the soluble fraction of clotted blood, is normally encountered by cells in vivo in the context of a wound. Indeed, the expression program that we observed in response to serum suggests that fibroblasts are programmed to interpret the abrupt exposure to serum not as a general mitogenic stimulus but as a specific physiological signal, signifying a wound. The proliferative response that we originally intended to study appeared to be part of a larger physiological response of fibroblasts to a wound. Other features of the transcriptional response to serum suggest that the fibroblast is an active participant in a conversation among the diverse cells that work together in wound repair, interpreting, amplifying, modifying, and broadcasting signals controlling inflammation, angiogenesis, and epithelial regrowth during the response to an injury.

We recognize that these in vitro results almost certainly represent a distorted and incomplete rendering of the normal physiological response of a fibroblast to a wound. Moreover, only the responses elicited directly by exposure of fibroblasts to serum were examined. The subsequent signals from other cellular participants in the normal wound-healing process would certainly provoke further evolution of the transcriptional program in fibroblasts at the site of a wound, which this experiment cannot reveal. Nevertheless, we believe that the picture that emerged strongly suggests a much larger and richer role for the fibroblast in the orchestration of this important physiological process than had previously been suspected.

References and Notes

1. J. A. Winkles, *Prog. Nucleic Acid Res. Mol. Biol.* **58**, 41 (1998).
2. A normal human diploid fibroblast cell line derived from foreskin (ATCC CRL 2091) in passage 8 was used in these experiments. The protocol followed for growth arrest and stimulation was essentially that of (*16*) and (*17*). Cells were grown to about 60% confluence in 15-cm petri dishes in Dulbecco's minimum essential medium containing glucose (1 g/liter), the antibiotics penicillin and streptomycin, and 10% (by vol) FBS (Hyclone) that had been previously heat inactivated at 56°C for 30 min. The cells were then washed three times with the same medium lacking FBS, and low-serum medium (0.1% FBS) was added to the plates. After a 48-hour incubation, the medium was replaced with fresh medium containing 10% FBS. mRNA was isolated from several plates of cells harvested before serum stimulation; this mRNA served as the serum-starved or time-zero reference sample. Cells were harvested from batches of plates at 11 subsequent intervals (15 min, 30 min, 1, 2, 4, 6, 8, 12, 16, 20, and 24 hours) after the addition of serum. mRNA was also isolated from exponentially growing fibroblasts (not subjected to serum starvation). mRNA was isolated with the FastTrack mRNA isolation kit (Invitrogen), which involves lysis of the cells on the plate. The growth medium was removed, and the cells were quickly washed with phosphate-buffered saline at room temperature. The lysis buffer was added to the plate, transferred to tubes, and frozen in liquid nitrogen. Subsequent steps were performed according to the kit manufacturer's protocols.
3. The National Center for Biotechnology Information maintains the UniGene database as a resource for partitioning human sequences contained in GenBank into clusters representing distinct transcripts or genes (*18*, *19*). At the time this work began, this database contained about 40,000 such clusters. We selected a subset of 10,000 of these UniGene clusters for inclusion on gene expression microarrays. UniGene clusters were included only if they contained at least one clone from the I.M.A.G.E. human cDNA collection (*20*), so that a physical clone could easily be obtained (all I.M.A.G.E. clones are available commercially from a number of vendors). We attempted to include as complete as possible a set of the "named" human genes (about 4000) and genes that appeared to be closely related to named genes in other organisms (about an additional 2000). The remaining 4000 clones were chosen from among the "anonymous" UniGene clusters on the basis of inclusion on the human transcript map (www.ncbi.nlm.nih.gov/SCIENCE96/) and the lack of apparent homology to any other genes in the selected set. A physical clone representing each of the selected genes was obtained from Research Genetics. This "10K set" is included in a more recent "15K set" described at www.nhgri.nih.gov/DIR/LCG/15K/HTML/p15Ktop.html. Of these clones, 472 are absent from the current edition of UniGene and were presumed to be distinct genes. The remainders represent 8141 distinct clusters, or human genes, in UniGene. These clones, thus presumed to represent 8613 different genes, were used to print microarrays according to methods described previously (*21*, *22*).
4. One microgram of mRNA was used for making fluorescently labeled cDNA probes for hybridizing to the microarrays, with the protocol described previously (*23*). mRNA from the large batch of serum-starved cells was used to make cDNA labeled with Cy3. The Cy3-labeled cDNA from this batch of serum-starved cells served as the common reference probe in all hybridizations. mRNA samples from cells harvested immediately before serum stimulation, at intervals after serum stimulation, and from exponentially growing cells were used to make cDNA labeled with Cy5. Ten micrograms of yeast tRNA, 10 μg of polydeoxyadenylic acid, and 20 μg of human CoT1 DNA (Gibco-BRL) were added to the mixture of labeled probes in a solution containing 3× standard saline citrate (SSC) and 0.3% SDS and allowed to prehybridize at room temperature for 30 min before the probe was added to the surface of the microarray. Hybridizations, washes, and fluorescent scans were performed as described previously (*23*, *24*). All measurements, totaling more than 180,000 differential expression measurements, were stored in a computer database for analysis and interpretation.
5. The nominal identities of a number of cDNAs (currently about 3750) on the microarray were verified by sequencing. The clones that were sequenced included many of the genes whose expression changed substantially upon serum stimulation, as well as a large number of genes whose expression did not change substantially in the course of this experiment. About 85% of the clones on the current version of this microarray that were checked by resequencing were correctly identified. In all the figures, gene names or EST numbers are given only for those genes on the microarrays whose identities were reconfirmed by resequencing. In the cases where a human gene has more than one name in the literature, we have tried to use the name that is most evocative of its presumed role in this context. The remainder of the clones have been assigned a temporary identification number (format: SID######) and a putative identity pending sequence verification. The correct identities of these genes will be posted at our Web site (genome-www.stanford.edu/serum) as they are confirmed by resequencing.
6. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
7. Genes were selected for this analysis if either (i) their expression level deviated from that in quiescent fibroblasts by at least a factor of 2.20 in at least two of the samples from serum-stimulated cells or (ii) the standard deviation for the set of 13 values of $\log_2$(expression ratio) measured for the gene in this time course exceeded 0.7. In addition, observations in which the pixel-by-pixel correlation coefficients for the Cy3 and Cy5 fluorescence signals measured in a given array element were less than 0.6 were excluded. This selection criteria yielded a computationally manageable number of genes while minimizing the number of genes that were included because of noise in the data.
8. A more complete analysis and interpretation of the results of this experiment, as well as a searchable database, can be found at genome-www.stanford.edu/serum
9. K. J. Livak, S. J. Flood, J. Marmaro, W. Giusti, K. Deetz, *PCR Methods. Appl.* **4**, 357 (1995).
10. The apparent dip in the profile of COX2 at the 2-hour time point in the microarray data appears to result from a localized area of low intensity on the corresponding array scan resulting in an underestimation of the expression ratio. The expression ratios measured for mast/stem cell growth factor receptor are somewhat lower in the microarray data. This discrepancy is probably a consequence of the conservative background subtraction method used for quantitating the signal intensities on the array scans (*23*). The sequences of the PCR primer pairs (5′ to 3′) that were used are as follows: COX2, CCGTGGCTCTCTTGGCAG and CTAAGTTCTTTAGCACTCCTTGGCA; IL-8, CGATGCTGTGGAGCTGTATC and CCATGGTTTCACCAAAGATG; mast/stem cell factor receptor, ACAGAAGCCCGTGGTAGACC and GAGGCTGGGAGGAGGAAG; B4-2, AAACCCCCCTCAGGAAAGAG and CCATGAACAAGCTGGCCAT; and actin, AGTACTCCGTGTGGATCGGC and GCTGATCCACATCTGCTGGA
11. V. R. Iyer *et al.*, unpublished data. The gene expression data for the early time points in the presence of cycloheximide will be available at our Web site (genome-www.stanford.edu/serum)
12. T. Hunter, *Cell* **80**, 225 (1995).
13. J. Leppaluoto and H. Ruskoaho, *Ann. Med.* **24**, 153 (1992).
14. A. C. Chang *et al.*, *Mol. Cell. Endocrinol.* **112**, 241 (1995).
15. K. L. Madsen *et al.*, *Am. J. Physiol.* **274**, G96 (1998).
16. W. Krek and J. A. DeCaprio, *Methods Enzymol.* **254**, 114 (1995).
17. R. A. Tobey, J. G. Valdez, H. A. Crissman, *Exp. Cell Res.* **179**, 400 (1988).
18. M. S. Boguski and G. D. Schuler, *Nature Genet.* **10**, 369 (1995).
19. G. D. Schuler, *J. Mol. Med.* **75**, 694 (1997).
20. G. Lennon, C. Auffray, M. Polymeropoulos, M. B. Soares, *Genomics* **33**, 151 (1996).
21. I.M.A.G.E. clones were amplified by PCR in 96-well format with amino-linked primers at the 5′ end. Purified PCR products were suspended at a concentration of ~0.5 mg/ml in 3× SSC, and ~5 ng of each product was arrayed onto coated glass by means of procedures similar to those described previously (*22*). A total of 9996 elements were arrayed onto an area of 1.8 cm by 1.8 cm with the elements spaced 175 μm apart. The microarrays were then postprocessed to fix the DNA to the glass surface before hybridization with a procedure similar to previously described methods (*22*).
22. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
23. J. L. DeRisi, V. R. Iyer, P. O. Brown, *ibid.* **278**, 680 (1997).
24. J. DeRisi *et al.*, *Nature Genet.* **14**, 457 (1996).
25. We thank E. Chung for help with sequencing, A. Alizadeh for help with sequence verification, K. Ranade for advice on the TaqMan assay, and J. DeRisi and other members of the P.O.B. and D.B. labs for discussions. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450) and the National Cancer Institute (NIH CA 77097). V.R.I. was supported in part by an Institutional Training Grant in Genome Sciences (T32 HG00044) from the NHGRI. M.B.E. is an Alfred E. Sloan Foundation Postdoctoral Fellow in Computational Molecular Biology, and D.T.R is a Walter and Idun Berry Fellow. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute.

13 August 1998; accepted 13 November 1998

Exhibit E of Iyer Declaration
with Response dated 4/23/04
In USSN: 10/049,742

Systematic variation in gene expression patterns in human cancer cell lines

Douglas T. Ross[1], Uwe Scherf[5], Michael B. Eisen[2], Charles M. Perou[2], Christian Rees[2], Paul Spellman[2], Vishwanath Iyer[1], Stefanie S. Jeffrey[3], Matt Van de Rijn[4], Mark Waltham[5], Alexander Pergamenschikov[2], Jeffrey C.F. Lee[6], Deval Lashkari[7], Dari Shalon[6], Timothy G. Myers[8], John N. Weinstein[5], David Botstein[2] & Patrick O. Brown[1,9]

We used cDNA microarrays to explore the variation in expression of approximately 8,000 unique genes among the 60 cell lines used in the National Cancer Institute's screen for anti-cancer drugs. Classification of the cell lines based solely on the observed patterns of gene expression revealed a correspondence to the ostensible origins of the tumours from which the cell lines were derived. The consistent relationship between the gene expression patterns and the tissue of origin allowed us to recognize outliers whose previous classification appeared incorrect. Specific features of the gene expression patterns appeared to be related to physiological properties of the cell lines, such as their doubling time in culture, drug metabolism or the interferon response. Comparison of gene expression patterns in the cell lines to those observed in normal breast tissue or in breast tumour specimens revealed features of the expression patterns in the tumours that had recognizable counterparts in specific cell lines, reflecting the tumour, stromal and inflammatory components of the tumour tissue. These results provided a novel molecular characterization of this important group of human cell lines and their relationships to tumours *in vivo*.

Introduction

Cell lines derived from human tumours have been extensively used as experimental models of neoplastic disease. Although such cell lines differ from both normal and cancerous tissue, the inaccessibility of human tumours and normal tissue makes it likely that such cell lines will continue to be used as experimental models for the foreseeable future. The National Cancer Institute's Developmental Therapeutics Program (DTP) has carried out intensive studies of 60 cancer cell lines (the NCI60) derived from tumours from a variety of tissues and organs[1–4]. The DTP has assessed many molecular features of the cells related to cancer and chemotherapeutic sensitivity, and has measured the sensitivities of these 60 cell lines to more than 70,000 different chemical compounds, including all common chemotherapeutics (http://dtp.nci.nih.gov). A previous analysis of these data revealed a connection between the pattern of activity of a drug and its method of action. In particular, there was a tendency for groups of drugs with similar patterns of activity to have related methods of action[3,5–7].

We used DNA microarrays to survey the variation in abundance of approximately 8,000 distinct human transcripts in these 60 cell lines. Because of the logical connection between the function of a gene and its pattern of expression, the correlation of gene expression patterns with the variation in the phenotype of the cell can begin the process by which the function of a gene can be inferred. Similarly, the patterns of expression of known genes can reveal novel phenotypic aspects of the cells and tissues studied[8–10]. Here we present an analysis of the observed patterns of gene expression and their relationship to phenotypic properties of the 60 cell lines. The accompanying report[11] explores the relationship between the gene expression patterns and the drug sensitivity profiles measured by the DTP. The assessment of gene expression patterns in a multitude of cell and tissue types, such as the diverse set of cell lines we studied here, under diverse conditions *in vitro* and *in vivo*, should lead to increasingly detailed maps of the human gene expression program and provide clues as to the physiological roles of uncharacterized genes[11–16]. The databases, plus tools for analysis and visualization of the data, are available (http://genome-www.stanford.edu/nci60 and http://discover.nci.nih.gov).

Results

We studied gene expression in the 60 cell lines using DNA microarrays prepared by robotically spotting 9,703 human cDNAs on glass microscope slides[17,18]. The cDNAs included approximately 8,000 different genes: approximately 3,700 represented previously characterized human proteins, an additional 1,900 had homologues in other organisms and the remaining 2,400 were identified only by ESTs. Due to ambiguity of the identity of the cDNA clones used in these studies, we estimated that approximately 80% of the genes in these experiments were correctly identified. The identities of approximately 3,000 cDNAs

Departments of [1]Biochemistry, [2]Genetics, [3]Surgery and [4]Pathology, Stanford University School of Medicine, Stanford, California, USA. [5]Laboratory of Molecular Pharmacology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA. [6]Incyte Pharmaceuticals, Fremont, California, USA. [7]Genometrix Inc., The Woodlands, Texas, USA. [8]Information Technology Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Rockville, Maryland, USA. [9]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California, USA. Correspondence should be addressed to P.O.B. (e-mail: pbrown@cmgm.stanford.edu) or J.N.W. (e-mail: Weinstein@dtpax2.ncifcrf.gov).

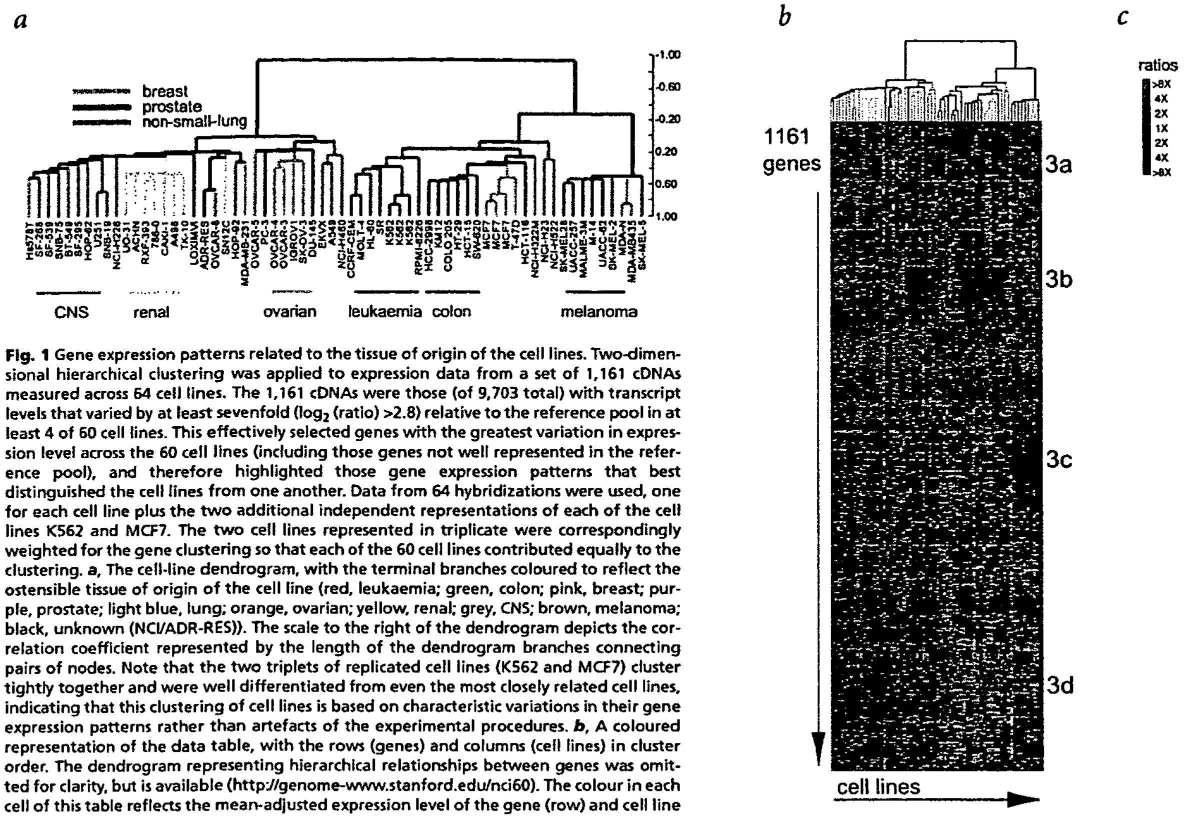


Fig. 1 Gene expression patterns related to the tissue of origin of the cell lines. Two-dimensional hierarchical clustering was applied to expression data from a set of 1,161 cDNAs measured across 64 cell lines. The 1,161 cDNAs were those (of 9,703 total) with transcript levels that varied by at least sevenfold ($\log_2$ (ratio) >2.8) relative to the reference pool in at least 4 of 60 cell lines. This effectively selected genes with the greatest variation in expression level across the 60 cell lines (including those genes not well represented in the reference pool), and therefore highlighted those gene expression patterns that best distinguished the cell lines from one another. Data from 64 hybridizations were used, one for each cell line plus the two additional independent representations of each of the cell lines K562 and MCF7. The two cell lines represented in triplicate were correspondingly weighted for the gene clustering so that each of the 60 cell lines contributed equally to the clustering. ***a***, The cell-line dendrogram, with the terminal branches coloured to reflect the ostensible tissue of origin of the cell line (red, leukaemia; green, colon; pink, breast; purple, prostate; light blue, lung; orange, ovarian; yellow, renal; grey, CNS; brown, melanoma; black, unknown (NCI/ADR-RES)). The scale to the right of the dendrogram depicts the correlation coefficient represented by the length of the dendrogram branches connecting pairs of nodes. Note that the two triplets of replicated cell lines (K562 and MCF7) cluster tightly together and were well differentiated from even the most closely related cell lines, indicating that this clustering of cell lines is based on characteristic variations in their gene expression patterns rather than artefacts of the experimental procedures. ***b***, A coloured representation of the data table, with the rows (genes) and columns (cell lines) in cluster order. The dendrogram representing hierarchical relationships between genes was omitted for clarity, but is available (http://genome-www.stanford.edu/nci60). The colour in each cell of this table reflects the mean-adjusted expression level of the gene (row) and cell line (column). The colour scale used to represent the expression ratios is shown. The labels '3a–3d' in (*b*) refer to the clusters of genes shown in detail in Fig. 3.

from these experiments have been sequence-verified, including all of those referred to here by name.

Each hybridization compared Cy5-labelled cDNA reverse transcribed from mRNA isolated from one of the cell lines with Cy3-labelled cDNA reverse transcribed from a reference mRNA sample. This reference sample, used in all hybridizations, was prepared by combining an equal mixture of mRNA from 12 of the cell lines (chosen to maximize diversity in gene expression as determined primarily from two-dimensional gel studies[2]). By comparing cDNA from each cell line with a common reference, variation in gene expression across the 60 cell lines could be inferred from the observed variation in the normalized Cy5/Cy3 ratios across the hybridizations.

To assess the contribution of artefactual sources of variation in the experimentally measured expression patterns, K562 and MCF7 cell lines were each grown in three independent cultures, and the entire process was carried out independently on mRNA extracted from each culture. The variance in the triplicate fluorescence ratio measurements approached a minimum when the fluorescence signal was greater than approximately 0.4% of the measurable total signal dynamic range above background in either channel of the hybridization. We selected the subset of spots for which significant signal was present in both the numerator and denominator of the ratios by this criterion to identify the best-measured spots. The pair-wise correlation coefficients for the triplicates of the set of genes that passed this quality control level (6,992 spots included for the MCF7 samples and 6,161 spots for K562) ranged from 0.83 to 0.92 (for graphs and details, see http://genome-www.stanford.edu/nci60).

To make the orderly features in the data more apparent, we used a hierarchical clustering algorithm[19,20] and a pseudo-colour visualization matrix[3,21]. The object of the clustering was to group cell lines with similar repertoires of expressed genes and to group genes whose expression level varied among the 60 cell lines in a similar manner. Clustering was performed twice using different subsets of genes to assess the robustness of the analysis. In one case (Fig. 1), we concentrated on those genes that showed the most variation in expression among the 60 cell lines (1,167 total). A second analysis (Fig. 2) included all spots that were thought to be well measured in the reference set (6,831 spots).

Gene expression patterns related to the histologic origins of the cell lines

The most notable property of the clustered data was that cell lines with common presumptive tissues of origin grouped together (Figs 1*a* and 2). Cell lines derived from leukaemia, melanoma, central nervous system, colon, renal and ovarian tissue were clustered into independent terminal branches specific to their respective organ types with few exceptions. Cell lines derived from non-small lung carcinoma and breast tumours were distributed in multiple different terminal branches suggesting that their gene expression patterns were more heterogeneous.

Many of these coherent cell line clusters were distinguished by the specific expression of characteristic groups of genes (Fig. 3*a*–*d*). For example, a cluster of approximately 90 genes was highly expressed in the melanoma-derived lines (Fig. 3*c*). This set was enriched for genes with known roles in melanocyte biology, including tyrosinase and dopachrome tautomerase (TYR and DCT; two subunits of an enzyme complex involved in melanin synthesis[22]), MART1 (MLANA; which is being investigated as a target for immunotherapy of melanoma[23]) and S100-β (S100B; which has been used as an antigenic marker in the diagnosis of

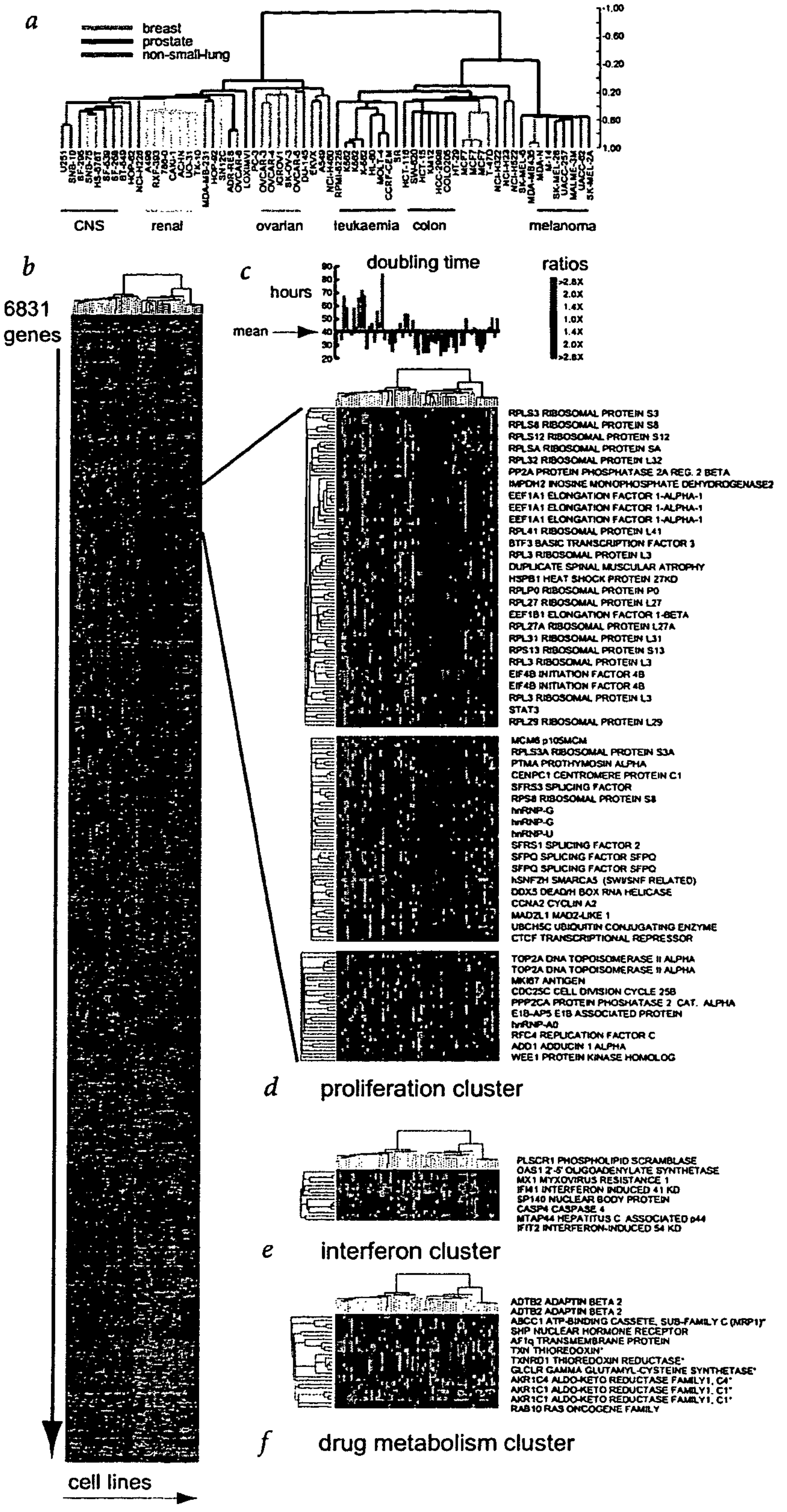


Fig. 2 Gene expression patterns related to other cell-line phenotypes. ***a***, We applied two-dimensional hierarchical clustering to expression data from a set of 6,831 cDNAs measured across the 64 cell lines. The 6,831 cDNAs were those with a minimum fluorescence signal intensity of approximately 0.4% of the dynamic range above background in the reference channel in each of the six hybridizations used to establish reproducibility. This effectively selected those spots that provided the most reliable ratio measurements and therefore identified a subset of genes useful for exploring patterns comprised of those whose variation in expression across the 60 cell lines was of moderate magnitude. ***b***, Cluster-ordered data table. ***c***, Doubling time of cell lines. Cell lines are given in cluster order. Values are plotted relative to the mean. Doubling times greater than the mean are shown in green, those with doubling time less than the mean are shown in red. ***d***, Three related gene clusters that were enriched for genes whose expression level variation was correlated with cell line proliferation rate. Each of the three gene clusters (clustered solely on the basis of their expression patterns) showed enrichment for sets of genes involved in distinct functional categories (for example, ribosomal genes versus genes involved in pre-RNA splicing). ***e***, Gene cluster in which all characterized and sequence-verified cDNAs encode genes known to be regulated by interferons. ***f***, Gene cluster enriched for genes that have been implicated in drug metabolism (indicated by asterisks). A further property of the gene clustering evident here and in Fig. 2 is the strong tendency for redundant representations of the same gene to cluster immediately adjacent to one another, even within larger groups of genes with very similar expression patterns. In addition to illustrating the reproducibility and consistency of the measurements, and providing independent confirmation of many of our measurements, this property also demonstrates that these, and probably all, genes have nearly unique patterns of variation across the 60 cell lines. If this were not the case, and multiple genes had identical patterns of variation, we would not expect to be able to distinguish, by clustering on the basis of expression variation, duplicate copies of individual genes from the other genes with identical expression patterns.

melanoma). LOXIMVI, the seventh line designated as melanoma in the NCI60, did not show this characteristic pattern. Although isolated from a patient with melanoma, LOXIMVI has previously been noted to lack melanin and other markers useful for identification of melanoma cells[1].

Paradoxically, two related cell lines (MDA-MB435 and MDA-N), which were derived from a single patient with breast cancer and have been conventionally regarded as breast cancer cell lines, shared expression of the genes associated with melanoma. MDA-MB435 was isolated from a pleural effusion in a patient with metastatic ductal adenocarcinoma of the breast[24,25]. It remains possible that the origin of the cell line was a breast cancer, and that its gene expression pattern is related to the neuroendocrine features of some breast cancers[26]. But our results suggest that this cell line may have originated from a melanoma, raising the possibility that the patient had a co-existing occult melanoma.

The higher-level organization of the cell-line tree—in which groups span cell lines from different tissue types—also reflected shared biological properties of the tissues from which the cell lines were derived. The carcinoma-derived cell lines were divided into major branches that separated those that expressed genes characteristic of epithelial cells from those that expressed genes more typical of stromal cells. A cluster of genes is shown (Fig. 3*b*) that is most strongly expressed in cell lines derived from colon carcinomas, six of seven ovarian-derived cell lines and the two breast cancer lines positive for the oestrogen receptor. The named genes in this cluster have been implicated in several aspects of epithelial cell biology[27]. The cluster was enriched for genes whose products are known to localize to the basolateral membrane of epithelial cells, including those encoding components of adherens complexes (for example, desmoplakin (DSP), periplakin (PPL) and plakoglobin (JUP)), an epithelial-expressed cell-cell adhesion molecule (M4S1) and a sodium/hydrogen ion exchanger[28–31] (SLC9A1). It also contained genes that encode putative transcriptional regulators of epithelial morphogenesis, a human homologue of a *Drosophila melanogaster* epithelial-expressed tumour suppressor (LLGL1) and a homeobox gene thought to control calcium-mediated adherence in epithelial cells[32,33] (MSX2).

In contrast, a separate, major branch of the cell-line dendrogram (Fig. 1*a*) included all glioblastoma-derived cell lines, all renal-cell-carcinoma–derived cell lines and the remaining carcinoma-derived lines. The characteristic set of genes expressed in this cluster included many whose products are involved in stromal cell functions (Fig. 3*d*). Indeed, the two cell lines originally described as 'sarcoma-like' in appearance (Hs578T, breast carcinosarcoma, and SF539, gliosarcoma) expressed most of these genes[34,35]. Although no single gene was uniformly characteristic of this cluster, each cell line showed a distinctive pattern of expression of genes encoding proteins with roles in synthesis or modification of the extracellular matrix (for example, caldesmon (CALD1), cathepsins, thrombospondin (THBS), lysyl oxidase (LOX) and collagen subtypes). Although the ovarian and most non-small-cell-lung–derived carcinomas expressed genes characteristic of both epithelial cells and stromal cells, they probably clustered with the CNS and renal cell carcinomas in this analysis because genes characteristically expressed in stromal cells were more abundantly represented in this gene set.

Physiological variation reflected in gene expression patterns

A cluster diagram of 6,831 genes (Fig. 2) is useful for exploring clusters of genes whose variation in mRNA levels was not obviously attributable to cell or tissue type. We identified some gene clusters that were enriched for genes involved in specific cellular processes; the variation in their expression levels may reflect corresponding differences in activity of these processes in the cell lines. For example, a cluster of 1,159 genes (Fig. 2*a*) included many whose products are necessary for progression through the cell cycle (such as CCNA1, MCM106 and MAD2L1), RNA processing and translation machinery (such as RNA helicases, hnRNPs and translation elongation factors) and traditional pathologic markers used to identify proliferating cells (MKI67). Within this large cluster were smaller clusters enriched for genes with more specialized roles. One cluster was highly enriched for numerous ribosomal genes, whereas another was more enriched for genes encoding RNA-splicing factors. The variation in expression of these ribosomal genes was significantly correlated with variation in the cell doubling time (correlation coefficient of 0.54), supporting the notion that the genes in this cluster were regulated in relation to cell proliferation rate or growth rate in these cell lines.

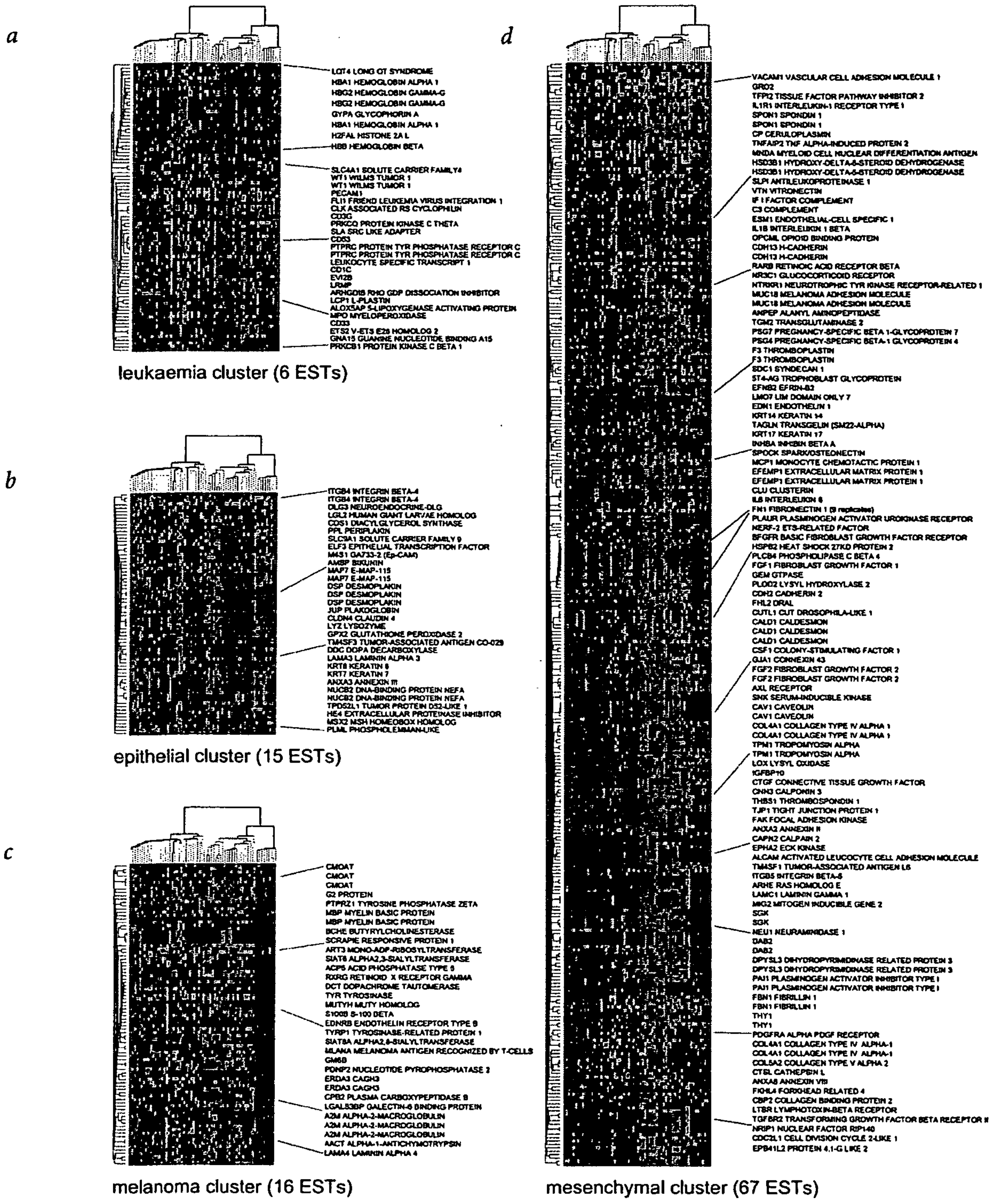
In a smaller gene cluster (Fig. 2*d*), all of the named genes were previously known to be regulated by interferons[13,36]. Additional groups of interferon-regulated genes showed distinct patterns of expression (data not shown), suggesting that the NCI60 cell lines exhibited variation in activity of interferon-response pathways, which was reflected in gene expression patterns[36].

Another cluster (Fig. 2*e*) contained several genes encoding proteins with possible interrelated roles in drug metabolism, including glutamate-cysteine ligase (GLCLC, the enzyme responsible for the rate limiting step of glutathione synthesis), thioredoxin (TXN) and thioredoxin reductase (TXNRD1; enzymes involved in regulating redox state in cells), and MRP1 (a drug transporter known to efficiently transport glutathione-conjugated compounds[37]). The elevated expression of this set of genes in a subset of these cell lines may reflect selection for resistance to chemotherapeutics.

Cell lines facilitate interpretation of gene expression patterns in complex clinical samples

Like many other types of cancer, tumours of the breast typically have a complex histological organization, with connective tissue and leukocytic infiltrates interwoven with tumour cells. To explore the possibility that variation in gene expression in the tumour cell lines might provide a framework for interpreting the expression patterns in tumour specimens, we compared RNA isolated from two breast cancer biopsy samples, a sample of normal breast tissue and the NCI60 cell lines derived from breast cancers (excluding MDA-MB-435 and MDA-N) and leukaemias (Fig. 4). This clustering highlighted features of the gene expression pattern shared between the cancer specimens and individual cell lines derived from breast cancers and leukaemias.

The genes encoding keratin 8 (KRT8) and keratin 19 (KRT19), as well as most of the other 'epithelial' genes defined in the complete NCI60 cell line cluster, were expressed in both of the biopsy samples and the two breast-derived cell lines, MCF-7 and T47D, expressing the oestrogen receptor, suggesting that these transcripts originated in tumour cells with features similar to those of luminal epithelial cells (Fig. 5*a*). Expression of a set of genes characteristic of stromal cells, including collagen genes (*COL3A1*, *COL5A1* and *COL6A1*) and smooth muscle cell markers (*TAGLN*), was a feature shared by the tumour sample and the stromal-like cell lines Hs578T and BT549 (Fig. 5*b*). This feature of the expression pattern seen in the tumour samples is likely to be due to the stromal component of the tumour. The tumours also shared expression of a set of genes (Fig. 5*c*) with the multiple myeloma cell line (RPMI-8226), notably including immunoglobulin genes, consistent with the presence of B cells in the tumour (this was confirmed by staining with anti-



leukaemia cluster (6 ESTs)

epithelial cluster (15 ESTs)

melanoma cluster (16 ESTs)

mesenchymal cluster (67 ESTs)

Fig. 3 Gene clusters related to tissue characteristics in the cell lines. Enlargements of the regions of the cluster diagram in Fig. 1 showing gene clusters enriched for genes expressed in cell lines of ostensibly similar origins. ***a***, Cluster of genes highly expressed in the leukaemia-derived cell lines. Two sub-clusters distinguish genes that were expressed in most leukaemia-derived lines from those expressed exclusively in the eryroblastoid line, K562 (note that the triplicate hybridizations cluster together). ***b***, Cluster of genes highly expressed in all colon (7/7) cell lines and all breast-derived cell lines positive for the oestrogen receptor (2/2). This set of genes was also moderately expressed in most ovarian lines (5/6) and some non-small-cell-lung (4/6) lines, but was expressed at a lower level in all renal-cancer-derived lines. ***c***, Cluster of genes highly expressed in most melanoma-derived lines (6/7) and two related lines ostensibly derived from breast cancer (MDA-MB435 and MDA-N). ***d***, Cluster of genes highly expressed in all glioblastoma (6/6) lines and most lines derived from renal-cell carcinoma (7/8), and more moderately expressed in a subset of carcinoma-derived lines. In all panels, names are shown only for all known genes whose identities were independently re-verified by sequencing. The number of sequence-validated ESTs within the cluster is indicated below the cluster in parentheses. The position of gene names in the adjacent list only approximates their position in the cluster diagram as indicated by the lines connecting the colour chart with the gene list. Complete cluster images with all gene names and accession numbers are available (http://genome-www.stanford.edu/nci60).

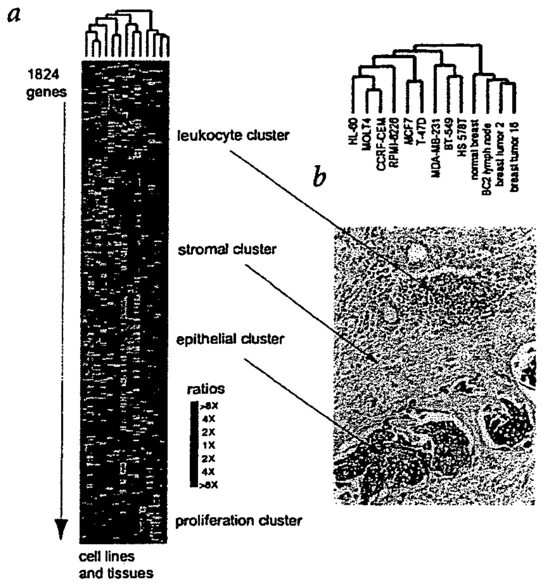


Fig. 4 Comparison of the gene expression patterns in clinical breast cancer specimens and cultured breast cancer and leukaemia cell lines. ***a***, Two-dimensional hierarchical clustering applied to gene expression data for two breast cancer specimens, a lymph node metastasis from one patient, normal breast and the NCI60 breast and leukaemia-derived cell lines. The gene expression data from tissue specimens was clustered along with expression data from a subset of the NCI60 cell lines to explore whether features of expression patterns observed in specific lines could be identified in the tissue samples. Labels indicate gene clusters (shown in detail in Fig. 5) that may be related to specific cellular components of the tumour specimens. ***b***, Breast cancer specimen 16 stained with anti-keratin antibodies, showing the complex mix of cell types characteristically found in breast tumours. The arrows highlight the different cellular components of this tissue specimen that were distinguished by the gene expression cluster analysis (Fig. 5).

immunoglobulin antibodies; data not shown). Therefore, distinct sets of genes with co-varying expression among the samples (Fig. 4, arrow) appear to represent distinct cell types that can be distinguished in breast cancer tissue. A fourth cluster of genes, more highly expressed in all of the cell lines than in any of the clinical specimens, was enriched for genes present in the 'proliferation' cluster described above (Fig. 5*d*). The variation in expression of these genes likely paralleled the difference in proliferation rate between the rapidly cycling cultured cell lines and the much more slowly dividing cells in tissues.

Discussion

Newly available genomics tools allowed us to explore variation in gene expression on a genomic scale in 60 cell lines derived from diverse tumour tissues. We used a simple cluster analysis to identify the prominent features in the gene expression patterns that appeared to reflect 'molecular signatures' of the tissue from which the cells originated. The histological characteristics of the cell lines that dominated the clustering were pervasive enough that similar relationships were revealed when alternative subsets of genes were selected for analysis. Additional features of the expression pattern may be related to variation in physiological attributes such as proliferation rate and activity of interferon-response pathways.

The properties of the tumour-derived cell lines in this study have presumably all been shaped by selection for resistance to host defences and chemotherapeutics and for rapid proliferation in the tissue culture environment of synthetic growth media, fetal bovine serum and a polystyrene substratum. But the primary identifiable factor accounting for variation in gene expression patterns among these 60 cell lines was the identity of the tissue from which each cell line was ostensibly derived. For most of the cell lines we examined, neither physiological nor experimental adaptation for growth in culture was sufficient to overwrite the gene expression programs established during differentiation *in vivo*. Nevertheless, the prominence of mesenchymal features in the cell lines isolated from glioblastomas and carcinomas may reflect a selection for the relative ease of establishment of cell lines expressing stromal characteristics, perhaps combined with physiological adaptation to tissue culture conditions[38–40].

Biological themes linking genes with related expression patterns may be inferred in many cases from the shared attributes of known genes within the clusters. Uncharacterized cDNAs are likely to encode proteins that have roles similar to those of the known gene products with which they appear to be co-regulated. Still, for several clusters of genes, we were unable to discern a common theme linking the identified members of the cluster. Further exploration of their variation in expression under more diverse conditions and more comprehensive investigation of the physiology of the NCI60 cells may provide insight[10]. The relationship of the gene expression patterns to the drug sensitivity patterns measured by the DTP is an example of linking variation in gene expression with more subtle and diverse phenotypic variation[11].

The patterns of gene expression measured in the NCI60 cell lines provide a framework that helps to distinguish the cells that express specific sets of genes in the histologically complex breast cancer specimens[41]. Although it is now feasible to analyse gene expression in micro-dissected tumour specimens[42,43], this observation suggests that it will be possible to explore and interpret some of the biology of clinical tumour samples by sampling them intact. As is useful in conventional morphological pathology, one might be able to observe interactions between a tumour and its microenvironment in this way. These relationships will be clarified by suitable analysis of gene expression patterns from intact as well as dissected tumours[12,14,15,41].

Methods

cDNA clones. We obtained the 9,703 human cDNA clones (Research Genetics) used in these experiments as bacterial colonies in 96-well microtitre plates[9]. Approximately 8,000 distinct Unigene clusters (representing nominally unique genes) were represented in this set of clones. All genes identified here by name represent clones whose identities were confirmed by re-sequencing, or by the criteria that two or more independent cDNA clones ostensibly representing the same gene had nearly identical gene expression patterns. A single-pass 3′ sequence re-verification was attempted for every clone after re-streaking for single colonies. For a subset of genes for which quality 3′ sequence was not obtained, we attempted to confirm identities by 5′ sequencing. Of the subset of clones selected for 5′ sequence verification on the basis of an interesting pattern of expression (888 total), 331 were correctly identified, 57, incorrectly identified, and 500, indeterminate (poor quality sequence). We estimated that 15%–20% of array elements contained DNA representing more than one clone per well. So far, the identities of ~3,000 clones have been verified. The full list of clones used and their nominal identities are available (gene names preceded by the designation "SID#" (Stanford Identification) represent clones whose identities have not yet been verified; http://genome-www.stanford.edu:8000/nci60).

Production of cDNA microarrays. The arrays used in this experiment were produced at Synteni Inc. (now Incyte Pharmaceuticals). Each insert was amplified from a bacterial colony by sampling 1 µl of bacterial media and performing PCR amplification of the insert using consensus primers for the three plasmids represented in the clone set (5′-TTGTAAAACGACG GCCAGTG-3′, 5′-CACACAGGAAACAGCTATG-3′). Each PCR product

(100 µl) was purified by gel exclusion, concentrated and resuspended in 3×SSC (10 µl). The PCR products were then printed on treated glass microscope slides using a robot with four printing tips. Detailed protocols for assembling and operating a microarray printer, and printing and experimental application of DNA microarrays are available (http://cmgm.stanford.edu/pbrown).

Preparation of mRNA and reference pool. Cell lines were grown from NCI DTP frozen stocks in RPMI-1640 supplemented with phenol red, glutamine (2 mM) and 5% fetal calf serum. To minimize the contribution of variations in culture conditions or cell density to differential gene expression, we grew each cell line to 80% confluence and isolated mRNA 24 h after transfer to fresh medium. The time between removal from the incubator and lysis of the cells in RNA stabilization buffer was minimized (<1 min). Cells were lysed in buffer containing guanidium isothiocyanate and total RNA was purified with the RNeasy purification kit (Qiagen). We purified mRNA as needed using a poly(A) purification kit (Oligotex, Qiagen) according to the manufacturer's instructions. Denaturing agarose gel electrophoresis assessed the integrity and relative contamination of mRNA with ribosomal RNA.

The breast tumours were surgically excised from patients and rapidly transported to the pathology laboratory, where samples for microarray analysis were quickly frozen in liquid nitrogen and stored at −80 °C until use. A frozen tumour specimen was removed from the freezer, cut into small pieces (~50–100 mg each), immediately placed into 10–12 ml of Trizol reagent (Gibco-BRL) and homogenized using a PowerGen 125 Tissue Homogenizer (Fisher Scientific), starting at 5,000 r.p.m. and gradually increasing to ~20,000 r.p.m. over a period of 30–60 s. We processed the Trizol/tumour homogenate as described in the Trizol protocol, including an initial step to remove fat. Once total RNA was obtained, we isolated mRNA with a FastTrack 2.0 kit (Invitrogen) using the manufacturer's protocol for isolating mRNA starting from total RNA. The normal breast samples were obtained from Clontech.

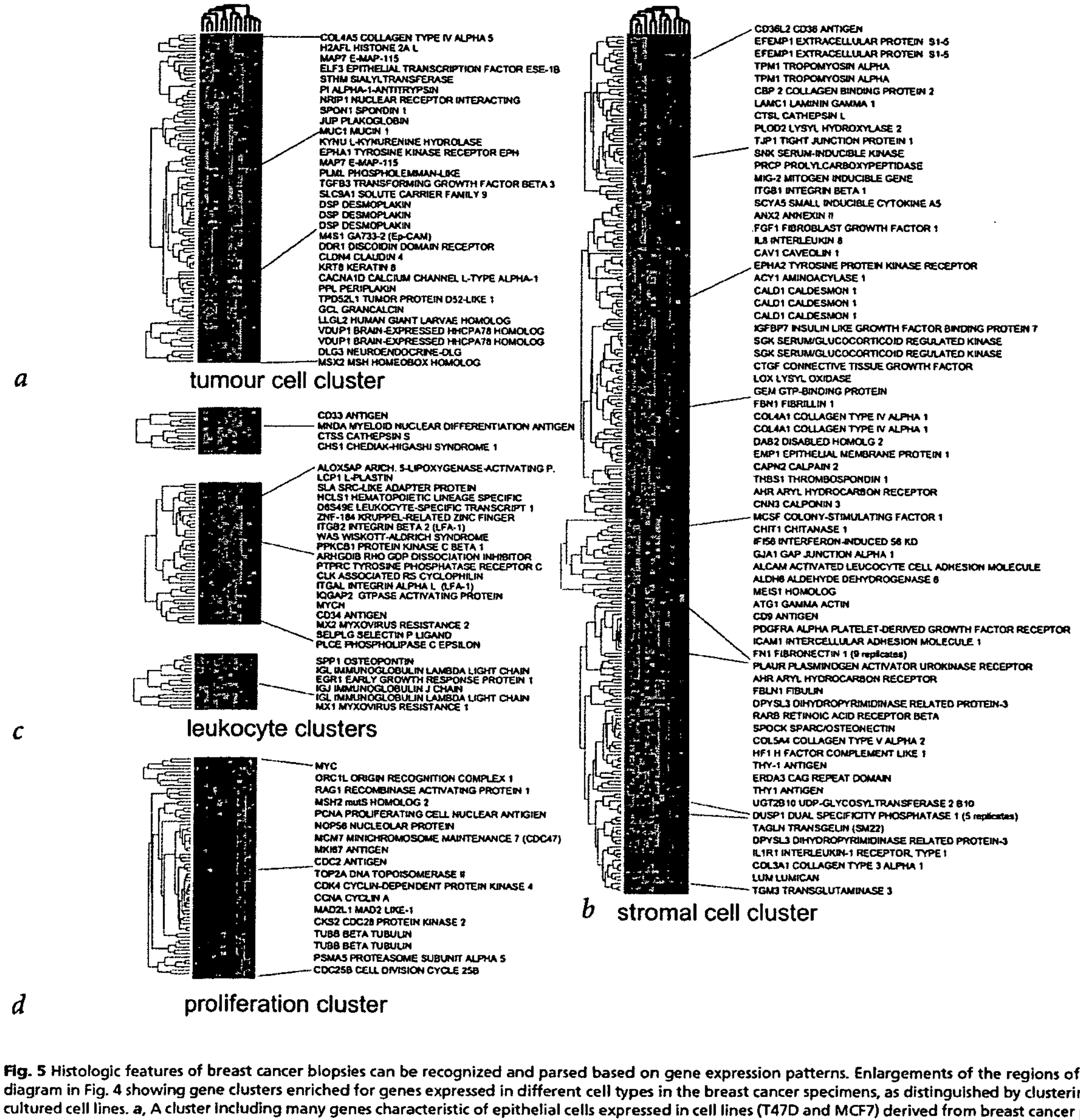


Fig. 5 Histologic features of breast cancer biopsies can be recognized and parsed based on gene expression patterns. Enlargements of the regions of the cluster diagram in Fig. 4 showing gene clusters enriched for genes expressed in different cell types in the breast cancer specimens, as distinguished by clustering with the cultured cell lines. ***a***, A cluster including many genes characteristic of epithelial cells expressed in cell lines (T47D and MCF7) derived from breast cancer positive for the oestrogen receptor and tumours. ***b***, Genes expressed in cell lines derived from breast cancer with stromal cell characteristics (Hs578T and BT549) and tumour specimens. Expression of these genes in the tumour samples may reflect the presence of myofibroblasts in the cancer specimen stroma. ***c***, Genes expressed in leukocyte-derived cell lines, showing common leukocyte, and separate 'myeloid' and 'B-cell', gene clusters. ***d***, Genes that were relatively highly expressed in all cell lines compared with the tumour specimens and normal breast. The higher expression of this set of genes involved in cell cycle transit in the cell lines is likely to reflect the higher proliferative rate of cells cultured in the presence of serum compared with the average proliferation rate of cells in the biopsied tissue.

We combined mRNA from the following cells in equal quantities to make the reference pool: HL-60 (acute myeloid leukaemia) and K562 (chronic myeloid leukaemia); NCI-H226 (non-small-cell-lung); COLO 205 (colon); SNB-19 (central nervous system); LOX-IMVI (melanoma); OVCAR-3 and OVCAR–4 (ovarian); CAKI-1 (renal); PC-3 (prostate); and MCF7 and Hs578T (breast). The criterion for selection of the cell lines in the reference are described in detail in the accompanying manuscript[12].

Doubling-time calculations. We calculated doubling times based on routine NCI60 cell line compound screening data; and they reflect the doubling times for cells inoculated into 96-well plates at the screening inoculation densities and grown in RPMI 1640 medium supplemented with 5% fetal bovine serum for 48 h. We measured cell populations using sulforhodamine B optical density measurement assay. The doubling time constant k was calculated using the equation: $N/No = e^{kt}$, where No is optical density for control (untreated) cells at time zero, *N* is optical density for control cells after 48-h incubation, and t is 48 h. The same equation was then used with the derived k to calculate the doubling time t by setting N/No = 2. For a given cell line, we obtained No and N values by averaging optical densities (N>6,000) obtained for each cell line for a year's screening. Data and experimental details are available (http://dtp.nci.nih.gov).

Preparation and hybridization of fluorescent labelled cDNA. For each comparative array hybridization, labelled cDNA was synthesized by reverse transcription from test cell mRNA in the presence of Cy5-dUTP, and from the reference mRNA with Cy3-dUTP, using the Superscript II reverse-transcription kit (Gibco-BRL). For each reverse transcription reaction, mRNA (2 µg) was mixed with an anchored oligo-dT (d-20T-d(AGC)) primer (4 µg) in a total volume of 15 µl, heated to 70 °C for 10 min and cooled on ice. To this sample, we added an unlabelled nucleotide pool (0.6 µl; 25 mM each dATP, dCTP, dGTP, and 15 mM dTTP), either Cy3 or Cy5 conjugated dUTP (3 µl; 1 mM; Amersham), 5×first-strand buffer (6 µl; 250 mM Tris-HCL, pH 8.3, 375 mM KCl, 15 mM $MgCl_2$), 0.1 M DTT (3 µl) and 2 µl of Superscript II reverse transcriptase (200 µ/µl). After a 2-h incubation at 42 °C, the RNA was degraded by adding 1 N NaOH (1.5 µl) and incubating at 70 °C for 10 min. The mixture was neutralized by adding of 1 N HCL (1.5 µl), and the volume brought to 500 µl with TE (10 mM Tris, 1 mM EDTA). We added Cot1 human DNA (20 µg; Gibco-BRL), and purified the probe by centrifugation in a Centricon-30 micro-concentrator (Amicon). The two separate probes were combined, brought to a volume of 500 µl, and concentrated again to a volume of less than 7 µl. We added 10 µg/µl poly(A) RNA (1 µl; Sigma) and tRNA (10 µg/µl; Gibco-BRL) were added, and adjusted the volume to 9.5 µl with distilled water. For final probe preparation, 20×SSC (2.1 µl; 1.5 M NaCl, 150 mM NaCitrate, pH 8.0) and 10% SDS (0.35 µl) were added to a total final volume of 12 µl. The probes were denatured by heating for 2 min at 100 °C, incubated at 37 °C for 20–30 min, and placed on the array under a 22 mm×22 mm glass coverslip. We incubated slides overnight at 65 °C for 14–18 h in a custom slide chamber with humidity maintained by a small reservoir of 3×SSC. Arrays were washed by submersion and agitation for 2–5 min in 2×SSC with 0.1% SDS, followed by 1×SSC and then 0.1×SSC. The arrays were "spun dry" by centrifugation for 2 min in a slide-rack in a Beckman GS-6 tabletop centrifuge in Microplus carriers at 650 r.p.m. for 2 min.

Array quantitation and data processing. Following hybridization, arrays were scanned using a laser-scanning microscope (ref. 17; http://cmgm.stanford.edu/pbrown). Separate images were acquired for Cy3 and Cy5. We carried out data reduction with the program ScanAlyze (M.B.E., available at http://rana.stanford.edu/software). Each spot was defined by manual positioning of a grid of circles over the array image. For each fluorescent image, the average pixel intensity within each circle was determined, and a local background was computed for each spot equal to the median pixel intensity in a square of 40 pixels in width and height centred on the spot centre, excluding all pixels within any defined spots. Net signal was determined by subtraction of this local background from the average intensity for each spot. Spots deemed unsuitable for accurate quantitation because of array artefacts were manually flagged and excluded from further analysis. Data files generated by ScanAlyze were entered into a custom database that maintains web-accessible files. Signal intensities between the two fluorescent images were normalized by applying a uniform scale factor to all intensities measured for the Cy5 channel. The normalization factor was chosen so that the mean log(Cy3/Cy5) for a subset of spots that achieved a minimum quality parameter (approximately 6,000 spots) was 0. This effectively defined the signal-intensity-weighted 'average' spot on each array to have a Cy3/Cy5 ratio of 1.0.

Cluster analysis. We extracted tables (rows of genes, columns of individual microarray hybridizations) of normalized fluorescence ratios from the database. Various selection criteria, discussed in relation to each data set, were applied to select subsets of genes from the 9,703 cDNA elements on the arrays. Before clustering and display, the logarithm of the measured fluorescence ratios for each gene were centred by subtracting the arithmetic mean of all ratios measured for that gene. The centring makes all subsequent analyses independent of the amount of each gene's mRNA in the reference pool.

We applied a hierarchical clustering algorithm separately to the cell lines and genes using the Pearson correlation coefficient as the measure of similarity and average linkage clustering[3,19–21]. The results of this process are two dendrograms (trees), one for the cell lines and one for the genes, in which very similar elements are connected by short branches, and longer branches join elements with diminishing degrees of similarity. For visual display the rows and columns in the initial data table were reordered to conform to the structures of the dendrograms obtained from the cluster analysis. Each cell in the cluster-ordered data table was replaced by a graded colour (pure red through black to pure green), representing the mean-adjusted ratio value in the cell. Gene labels in cluster diagrams are displayed here only for genes that were represented in the microarray by sequence-verified cDNAs. A complete software implementation of this process is available (http://rana.stanford.edu/software), as well as all clustering results (http://genome-www.stanford.edu/nci60).

Acknowledgements

We thank members of the Brown and Botstein labs for helpful discussions. This work was supported by the Howard Hughes Medical Institute and a grant from the National Cancer Institute (CA 077097). The work of U.S. and J.N.W. was supported in part by a grant from the National Cancer Institute Breast Cancer Think Tank. D.T.R is a Walter and Idun Berry Fellow. M.B.E. is an Alfred P. Sloan Foundation Fellow in Computational Molecular Biology. C.M.P. is a SmithKline Beecham Pharmaceuticals Fellow of the Life Science Research Foundation. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute.

Received 20 July 1999; accepted 13 January 2000.

1. Stinson, S.F. *et al.* Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. *Anticancer Res.* **12**, 1035–1053 (1992).
2. Myers, T.G. *et al.* A protein expression database for the molecular pharmacology of cancer. *Electrophoresis* **18**, 647–653 (1997).
3. Weinstein, J.N. *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
4. Monks, A., Scudiero, D.A., Johnson, G.S., Paull, K.D. & Sausville, E.A. The NCI anti-cancer drug screen: a smart screen to identify effectors of novel targets. *Anticancer Drug Des.* **12**, 533–541 (1997).
5. Paull, K.D. *et al.* Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl Cancer Inst.* **81**, 1088–1092 (1989).
6. Weinstein, J.N. *et al.* Neural computing in cancer drug development: predicting mechanism of action. *Science* **258**, 447–451 (1992).
7. van Osdol, W.W., Myers, T.G., Paull, K.D., Kohn, K.W. & Weinstein, J.N. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J. Natl Cancer Inst.* **86**, 1853–1859 (1994).
8. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
9. Iyer, V.R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87 (1999).
10. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21** (suppl.), 33–37 (1999).
11. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **24**, 236–244 (2000).
12. Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
13. Der, S.D., Zhou, A., Williams, B.R. & Silverman, R.H. Identification of genes differentially regulated by interferon-α, -β or -γ or using oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **95**, 15623–15628 (1998).
14. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
15. Wang, K. *et al.* Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* **229**, 101–108 (1999).
16. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
17. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645 (1996).
18. Eisen, M.B. & Brown, P.O. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303**, 179–205 (1999).
19. Sokal, R.R. & Sneath, P.H.A. *Principles of Numerical Taxonomy* (W.H. Freeman, San Francisco, 1963).
20. Hartigan, J.A. *Clustering Algorithms* (Wiley, New York, 1975).
21. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
22. del Marmol, V. & Beermann, F. Tyrosinase and related proteins in mammalian pigmentation. *FEBS Lett.* **381**, 165–168 (1996).
23. Kawakami, Y. *et al.* The use of melanosomal proteins in the immunotherapy of melanoma. *J. Immunother.* **21**, 237–246 (1998).
24. Cailleau, R., Olive, M. & Cruciger, Q.V. Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. *In Vitro* **14**, 911–915 (1978).
25. Brinkley, B.R. *et al.* Variations in cell form and cytoskeleton in human breast carcinoma cells in vitro. *Cancer Res.* **40**, 3118–3129 (1980).
26. Nesland, J.M., Holm, R., Johannessen, J.V. & Gould, V.E. Neuroendocrine differentiation in breast lesions. *Pathol. Res. Pract.* **183**, 214–221 (1988).
27. Davies, J.A. & Garrod, D.R. Molecular aspects of the epithelial phenotype. *Bioessays* **19**, 699–704 (1997).
28. Garrod, D., Chidgey, M. & North, A. Desmosomes: differentiation, development, dynamics and disease. *Curr. Opin. Cell Biol.* **8**, 670–678 (1996).
29. Cowin, P. & Burke, B. Cytoskeleton-membrane interactions. *Curr. Opin. Cell Biol.* **8**, 56–65 (1996); erratum: **8**, 244 (1996).
30. Litvinov, S.V. *et al.* Epithelial cell adhesion molecule (Ep-CAM) modulates cell-cell interactions mediated by classic cadherins. *J. Cell Biol.* **139**, 1337–1348 (1997).
31. Helmle-Kolb, C. *et al.* Na/H exchange activities in NHE1-transfected OK-cells: cell polarity and regulation. *Pflugers Arch.* **425**, 34–40 (1993); erratum: **427**, 387 (1994).
32. Manfruelli, P., Arquier, N., Hanratty, W.P. & Semeriva, M. The tumor suppressor gene, lethal(2)giant larvae (1(2)g1), is required for cell shape change of epithelial cells during Drosophila development. *Development* **122**, 2283–2294 (1996).
33. Lincecum, J.M., Fannon, A., Song, K., Wang, Y. & Sassoon, D.A. Msh homeobox genes regulate cadherin-mediated cell adhesion and cell-cell sorting. *J. Cell Biochem.* **70**, 22–28 (1998).
34. Hackett, A.J. *et al.* Two syngeneic cell lines from human breast tissue: the aneuploid mammary epithelial (Hs578T) and the diploid myoepithelial (Hs578Bst) cell lines. *J. Natl Cancer Inst.* **58**, 1795–1806 (1977).
35. Rutka, J.T. *et al.* Establishment and characterization of a cell line from a human gliosarcoma. *Cancer Res.* **46**, 5893–5902 (1986).
36. Nguyen, H., Hiscott, J. & Pitha, P.M. The growing family of interferon regulatory factors. *Cytokine Growth Factor Rev.* **8**, 293–312 (1997).
37. Moscow, J.A., Schneider, E., Ivy, S.P. & Cowan, K.H. Multidrug resistance. *Cancer Chemother. Biol. Response Modif.* **17**, 139–177 (1997).
38. Smith, H.S. & Hackett, A.J. The use of cultured human mammary epithelial cells in defining malignant progression. *Ann. N Y Acad. Sci.* **464**, 288–300 (1986).
39. Rutka, J.T. *et al.* Establishment and characterization of five cell lines derived from human malignant gliomas. *Acta Neuropathol.* **75**, 92–103 (1987).
40. Ronnov-Jessen, L., Petersen, O.W. & Bissell, M.J. Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction. *Physiol. Rev.* **76**, 69–125 (1996).
41. Perou, C.M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
42. Bonner, R.F. *et al.* Laser capture microdissection: molecular analysis of tissue. *Science* **278**, 1481–1483 (1997).
43. Sgroi, D.C. *et al.* In vivo gene expression profile analysis of human breast cancer progression. *Cancer Res.* **59**, 5656–5661 (1999).

PHYSIOLOGICAL REVIEWS
Vol. 79, No. 2, April 1999
Printed in U.S.A.

Chaperone-Mediated Protein Folding

ANTHONY L. FINK

Department of Chemistry and Biochemistry, The University of California, Santa Cruz, California

Fink, Anthony L. Chaperone-Mediated Protein Folding. *Physiol. Rev.* 79: 425–449, 1999.–The folding of most newly synthesized proteins in the cell requires the interaction of a variety of protein cofactors known as molecular chaperones. These molecules recognize and bind to nascent polypeptide chains and partially folded intermediates of proteins, preventing their aggregation and misfolding. There are several families of chaperones; those most involved in protein folding are the 40-kDa heat shock protein (HSP40; DnaJ), 60-kDa heat shock protein (HSP60; GroEL), and 70-kDa heat shock protein (HSP70; DnaK) families. The availability of high-resolution structures has facilitated a more detailed understanding of the complex chaperone machinery and mechanisms, including the ATP-dependent reaction cycles of the GroEL and HSP70 chaperones. For both of these chaperones, the binding of ATP triggers a critical conformational change leading to release of the bound substrate protein. Whereas the main role of the HSP70/HSP40 chaperone system is to minimize aggregation of newly synthesized proteins, the HSP60 chaperones also facilitate the actual folding process by providing a secluded environment for individual folding molecules and may also promote the unfolding and refolding of misfolded intermediates.

I. INTRODUCTION

The basic paradigm of molecular chaperones is that they recognize and selectively bind nonnative, but not native, proteins to form relatively stable complexes (48). In most cases, the complexes are dissociated by the binding and hydrolysis of ATP. In addition, there are "specific" molecular chaperones that typically are involved in the assembly of particular multiprotein complexes. Molecular chaperones comprise several highly conserved families of unrelated proteins; many chaperones are also heat shock (stress) proteins. The ubiquitous role of molecular chaperones continues to unfold with more discoveries each year. In the context of in vivo protein folding, chaperones prevent irreversible aggregation of nonnative conformations and keep proteins on the productive folding pathway. In addition, they may maintain newly synthesized proteins in an unfolded conformation suitable for trans-

0031-9333/99 $15.00 Copyright © 1999 the American Physiological Society

location across membranes and bind to nonnative proteins during cellular stress, among other functions. It is likely that most, if not all, cellular proteins will interact with a chaperone at some stage of their lifetime.

The focus of this review is on the functional contribution of chaperones to in vivo protein folding and assembly, especially those chaperones that are promiscuous, in that they show broad specificity for binding nonnative proteins. In addition to several specialized review articles on chaperones, e.g., References 11, 49, 50, 53, 76, 77, 90, 144, 164, 190, 191, 226, 228, 252, there have been two recent monographs published on the subject (60, 150). This review is of necessity selective; the main goal is to furnish an up-to-date overview of the role of the major molecular chaperones involved in protein folding. In view of the vast array of literature on molecular chaperones, and the ease of access to literature citations using the Internet, this article should not be viewed as exhaustive.

Why do we need chaperones? After all, a basic tenet of in vitro protein folding has been the seminal work of Anfinsen (2), which demonstrated that formation of the native protein from the unfolded state is a spontaneous process determined by the global free energy minimum. The results indicated that the native state of small globular proteins is determined by their amino acid sequence. However, the experimental conditions necessary to successfully fold many proteins, especially larger ones, in vitro, are very constrictive, usually requiring very low protein concentration and long incubation times and are usually unphysiological (e.g., relatively low temperatures). In contrast, most cells operate at ambient or homeothermically set temperatures (e.g., 37°C) where the hydrophobic effect will be stronger and thus protein denaturation and aggregation will be bigger problems, and the time-frame available for successful folding is short. Thus there is the need for additional factors for the successful folding of many proteins in vivo. When one considers the crowded cellular environment within a cell, it becomes clear that in vitro folding experiments at low protein concentrations are poor models for what happens in the cell, where a newly synthesized protein is in an environment with little or no "free" water, very high concentrations of other proteins and metabolites, and typically membranes, cytoskeletal elements, and other cellular components. Thus the need for chaperones *1*) to prevent aggregation and misfolding during the folding of newly synthesized chains, *2*) to prevent nonproductive interactions with other cell components, *3*) to direct the assembly of larger proteins and multiprotein complexes, and *4*) during exposure to stresses that cause previously folded proteins to unfold, becomes evident. In the few cases where folding has been studied both in vivo and in vitro, it appears that the folding pathways are similar (148, 190).

Cells have solved the problem of misfolding and aggregation, to a considerable extent at least, through the participation of molecular chaperones in the in vivo folding process. Many investigations in the past few years have confirmed the critical role of molecular chaperones in protein folding in the cell. Although much has been learned about the function of chaperones in protein folding, and the general outline of the process is thought to be understood, there are still many important unresolved issues, and new chaperones and cochaperones are still being discovered.

The molecular chaperones involved in the folding of newly synthesized proteins recognize nonnative substrate proteins predominantly via their exposed hydrophobic residues. The major chaperone classes are 40-kDa heat shock protein (HSP40; the DnaJ family), 60-kDa heat shock protein [HSP60; including GroEL and the T-complex polypeptide 1 (TCP-1) ring complexes], 70-kDa heat shock protein (HSP70), and 90-kDa heat shock protein (HSP90). All these chaperones can prevent the aggregation of at least some unfolded proteins. For HSP60 and HSP70, their activity is modulated by the binding and hydrolysis of ATP. The HSP70 (DnaK in *Escherichia coli*) bind to nascent polypeptide chains on ribosomes, preventing their premature folding, misfolding, or aggregation, as well as to newly synthesized proteins in the process of translocation from the cytosol into the mitochondria and the endoplasmic reticulum (ER). The HSP70 are regulated by HSP40 (DnaJ or its homologs). The HSP60 are large oligomeric ring-shaped proteins known as chaperonins that bind partially folded intermediates, preventing their aggregation, and facilitating their folding and assembly. This family is composed of GroEL-like proteins in eubacteria, mitochondria, and chloroplasts and the TCP-1 (CCT or TRiC) family in the eukaryotic cytosol and the archaea. The HSP60 (GroEL in *E. coli*) are large, usually tetradecameric proteins with a central cavity in which nonnative protein structures bind. The HSP60 are found in all biological compartments except the ER. The HSP60 are regulated by a cochaperone, chaperonin 10 (cpn10) (GroES in *E. coli*). In addition to preventing aggregation, it has been suggested that HSP60 may permit misfolded structures to unfold and refold. The HSP90 are associated with a number of proteins and play important roles in modulating their activity, most notably the steroid receptors. A number of other proteins involved in the folding of many newly synthesized proteins are often considered to be molecular chaperones; these include protein disulfide isomerase and peptidyl prolyl isomerase, which catalyze the rearrangement of disulfide bonds and isomerization of peptide bonds around Pro residues, respectively, and are perhaps better considered to be folding catalysts rather than chaperones. As mentioned previously, there are also a number of more specific chaperones that are involved in the folding/assembly of only one, or a very limited number, of particular substrate proteins.

Chaperones are catalysts in the sense that they transiently interact with their substrate proteins but are not present in the final folded product, and also in that they increase the yield of folded protein. However, there is no good evidence that they actually enhance the spontaneous rate of folding itself, although they may appear to do this by minimizing off-pathway reactions.

A brief perusal of the literature demonstrates that our knowledge of molecular chaperones is growing at an enormous pace. To put these new discoveries in context, a few more general points are worthy of note. Although in many respects the field of molecular chaperones can now be considered a mature one, in that it has passed its first decade of life, and the broad outlines, at least, are reasonably well established, there are still many outstanding questions. Furthermore, there are many areas of considerable controversy, and many of these relate to fundamental questions. For example, we do not yet know with certainty whether all newly synthesized proteins interact with chaperones, although it is likely that they do. We certainly do not know much about all the interactions between the various chaperones themselves, as well as with newly synthesized proteins or other chaperone target proteins. As discussed in this review, there are significant controversies concerning which chaperones interact first with nascent polypeptides, and even whether all nascent polypeptides interact with chaperones. The GroEL family of chaperones has been intensively studied, especially in the context of in vitro protein folding, yet it is not clear just how important a role this family (the cpn60 chaperonins and their TCP-1 eukaryotic homologs) play in the folding of most proteins in the cell. We are only now beginning to get a picture of the apparently ubiquitous role of the HSP90 family in many critical processes in the cell, especially those involving protein-protein interactions. Recently, several new "accessory" proteins have been discovered, which apparently act as "cochaperones." Again, their significance to protein folding and denaturation in the cell in general is unclear at the present time; they may be highly specialized or may turn out to be critical in a broad range of cellular processes involving chaperones. Although some of the chaperones clearly are important in preventing protein aggregation, there is as yet no good evidence that chaperones play a role significant in the opposite side of this equation, namely, in solubilizing protein aggregates, although it would seem likely that this may in fact be a function of some chaperones. Even at the level of the specific mechanisms of chaperone function, there are many controversial aspects, and those in the field know there have been some quite rancorous discussions over competing mechanisms. Thus the molecular chaperone field is one in which there are still many outstanding questions, including some quite fundamental ones. Consequently, chaperone scientists are likely to remain busy for a long time to come.

We begin with a brief review of the current understanding of in vitro protein folding and the potential for aggregation and misfolding.

II. IN VITRO PROTEIN FOLDING

Despite the fact that in vitro folding may not exactly mimic folding in the cell, it is minimally a good model for in vivo protein folding and has the critical advantage that a very wide variety of biophysical methods may be applied to provide a detailed knowledge of the folding pathway, kinetics, and energetics. Significant increases in our understanding of the folding process have occurred in the past few years, especially through the application of sophisticated new techniques, and these have been summarized in recent reviews (28, 29, 43, 44, 51, 56, 57, 177, 186, 255, 267). Both in vivo and in vitro, proteins fold remarkably rapidly, indicating that the folding pathway is directed in some way. Many studies have revealed intermediates during in vitro protein folding experiments; it is not clear, and is very difficult to establish experimentally, whether these are on- or off-pathway species. Although it is becoming apparent that in some cases these may be off-pathway species (208), some appear to be true intermediates on the productive folding pathway, consistent with rugged energy landscapes (248).

Small proteins may, under appropriate conditions, fold to the native state within a few tens of milliseconds with no detectable intermediates (177, 210). Such folding is consistent with smooth funnel energy landscape models (248), i.e., no intermediates, but could also reflect very fast folding with intermediates of sufficiently short lifetimes that they are not detected by current methods (44). However, for many systems there is substantial experimental data to support the presence of partially folded intermediates during folding. Although stopped-flow circular dichroism kinetics investigations reveal substantial secondary structure formation within a few milliseconds of the initiation of folding, most proteins take much longer to achieve the native state (seconds or longer).

The earliest stages of folding involve hydrophobic collapse to a relatively compact state and formation of metastable secondary structure. It is not clear if collapse or secondary structure occur simultaneously or if one precedes the other. It is most likely that both proceed concurrently. Certainly secondary structural units may be formed on a microsecond time scale (24). There is no conclusive data yet available on how fast the collapse occurs.

The nature of this initial collapsed state will vary depending on the conditions and the particular protein, but in general, it will consist of a very large number of substates. Further condensation will lead to one or more particularly stable intermediates; again depending on the particular protein, the intermediate(s) will have regions of unique structure, especially in terms of the compactness, amount of secondary structure, and topology. It is very likely that in most cases these intermediates will consist of a core of nativelike structure with the remainder of the protein in varying degrees of disorder. Regions of the nonordered chain are probably flickering in and out of their nativelike secondary structure conformation. At least some proteins fold via a hierarchical path in which additional structural units coalesce to an initially formed core with nativelike structure (61).

It is now clear, based on investigations of transient and equilibrium intermediates in vitro, that partially folded intermediates, as found with newly synthesized proteins in the cell, are particularly prone to aggregate, probably via specific intermolecular interactions between hydrophobic surfaces of structural subunits (59, 255). The intermediates are more prone to aggregate than the unfolded state because in the latter the hydrophobic side chains are scattered relatively randomly in many small hydrophobic regions, whereas in the partially folded intermediates, there will be large patches of contiguous surface hydrophobicity that will have a much stronger propensity for aggregation. The tendency of partially folded intermediates to associate or aggregate is exacerbated as the protein concentration increases. The growing recognition of the critical importance of protein aggregation has resulted in a number of reviews (42, 59, 112, 253–255).

A. Molecular Chaperones and Protein Aggregation

Both in vivo and in vitro the transition of a protein from the unfolded to folded state frequently results in the formation of partially folded intermediate states that have a very strong propensity to aggregate. In vivo this may lead to formation of inclusion bodies, especially when overexpression occurs. Members of the HSP60 and HSP70 molecular chaperone families seem to be most directly, and most generally, involved in preventing this. Current understanding of the role of HSP70 in protein folding suggests that the chaperone sequesters the unfolded or partially folded protein, thereby preventing its aggregation, but does not actively participate in the folding process; subsequent binding of ATP leads to release of the substrate protein in a nonnative conformation (144, 146, 166, 167). The *E. coli* HSP60 chaperone GroEL and its eukaryotic homologs facilitate protein folding by binding partially folded intermediates (or partially folded domains of large multidomain proteins) in their large central cavity (see sect. *vB*). Folding can thus occur in a situation where aggregation is precluded (144, 229). The general outline is summarized in Figure 1.

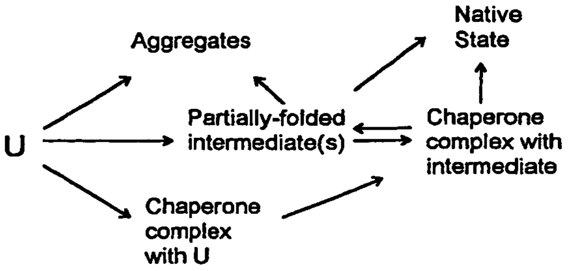

Aggregates
Native State
U
Partially-folded intermediate(s)
Chaperone complex with intermediate
Chaperone complex with U

FIG. 1. General outline of chaperone-mediated protein folding in a cell. U represents nascent polypeptide or newly synthesized protein. Chaperones include 40-, 60-, and 70-kDa heat shock proteins, protein disulfide isomerase, and peptidyl prolyl isomerase or trigger factor. For multisubunit proteins, the situation is more complex and not yet well understood.

It is likely that a significant factor in the formation of in vivo aggregates, such as inclusion bodies, is a lack of available molecular chaperones, usually due to either the rapid rate of protein synthesis, the formation of long-lived folding intermediates, or a combination of both. Either situation could lead to saturation of the available chaperones. The longer a protein takes to fold spontaneously, the longer it is likely to remain associated with the HSP60 and HSP70 chaperones. Some proteins fold fast and may have partially folded intermediates that have little propensity to aggregate, thus requiring little or no chaperone assistance and little tendency to form inclusion bodies.

Several experiments have been conducted in which overexpression of various combinations of the DnaK and GroEL chaperone systems decreases the amount of aggregation (8, 72, 81, 134, 229). For example, newly synthesized proteins in *E. coli* were shown to aggregate extensively when the *rpoH* mutation was present (81). This mutation in the RNA polymerase σ^{32}-subunit, which is responsible for heat shock promotor recognition, leads to a lack of heat shock proteins. Although growth is normal at 30°C, on elevating the temperature to 42°C, the cell is unable to produce sufficient chaperones and massive aggregation is observed. Overproduction of either GroEL and GroES, or DnaK and DnaJ, significantly decreases the aggregation at 42°C. If overexpressed together, the four chaperones are able to suppress most of the aggregation. The data suggest that the GroEL/GroES and the DnaK/DnaJ chaperone systems have complementary functions in the folding and assembly of most proteins. In addition, for in vitro aggregating systems, the presence of various chaperones increases the yield of soluble or native protein (16, 18, 97, 221). There have been conflicting reports as to whether the DnaK or GroEL systems, individually or together, yield the optimal amount of renaturation. It appears that in some cases all the chaperones are required for maximal suppression of aggregation, whereas in others either the DnaK system alone, or the GroEL system alone, was effective (229). It is possible that the two systems interact at different stages of folding, and thus different results may be observed depending on the particular system (16).

III. MOLECULAR CHAPERONES INVOLVED IN IN VIVO PROTEIN FOLDING

The major classes of general chaperones are the HSP40, HSP60, HSP70, HSP90, 100-kDa heat shock protein (HSP100), and the small heat shock proteins. Recent investigations have shown that not only do the major classes of chaperones often function with protein cofactors, but direct interactions between members of the HSP40, HSP70, and HSP90 families may be frequent. This section provides a brief description of the main families of molecular chaperones involved in protein folding in the cell.

A. Small Heat Shock Proteins and α-Crystallins

The small heat shock protein (HSP) and α-crystallin family consists of 12- to 43-kDa proteins that assemble into large multimeric structures and contain a conserved COOH-terminal region termed the α-crystallin domain. Many of the small HSP are produced only under stress conditions. They have been shown to function in vitro as chaperones by preventing protein aggregation in an ATP-independent manner. Several recent reviews have been published (19, 47, 113, 197). The role of small HSP in protein folding in vivo is unclear, but it seems unlikely that they are major players; this probably reflects the fact that release of bound, denatured proteins from the small heat shock proteins is very slow or nonexistent. For the α-crystallins in the eye lens, a major role is to bind denatured proteins and prevent their aggregation (which would result in cataracts). The small HSP bind denatured proteins tightly, but there is little evidence at present that they normally release the bound material subsequently. It has been proposed that their major function may be in times of stress when they bind denatured proteins and prevent their aggregation. Subsequently, when the stress is removed, these complexes may provide a reservoir for the HSP70 chaperone machinery to renature the bound proteins (47). The small HSP exhibit high affinity for partially folded intermediates but show no apparent substrate specificity and are only functional in the oligomeric form (131). Little is known about the mechanism of action of the small HSP; it has been suggested that the substrate protein coats the outside of the large chaperone multimer (133) and that hydrophobic interactions are critical in substrate binding. Several models have been proposed for the quaternary structure of the small HSP, but no consensus exists. A model in which small HSP prevent protein aggregation and may facilitate substrate refolding in conjunction with other molecular chaperones has recently been proposed (133).

B. HSP40 Family

The HSP40 or DnaJ family consists of over 100 members, defined by the presence of a highly conserved J domain of ~78 residues (DnaJ from *E. coli* has 376 amino acids) (131). Proteins in this family typically consist of several domains, e.g., DnaJ contains at least four conserved regions representing potential functional domains (the J domain, which is linked by a Gly/Phe-rich region to a domain of unknown function, followed by a zinc-finger region, and ending with the COOH-terminal domain, also of unknown function). Much variability is seen in the non-J domains of members of this family. The best studied examples are DnaJ from *E. coli* and several homologs from yeast, such as Mdj1 and Ydj1 (33, 34, 189). The best defined role thus far for the HSP40 is as a cochaperone for HSP70; however, even this function is not well understood, and there is evidence to indicate that DnaJ and other members of the HSP40 family are chaperones in their own right, binding to at least some unfolded proteins and nascent chains (94). The details of the putative role of DnaJ in protein folding are described in sections IV and V. In *E. coli*, DnaK, DnaJ, and GrpE cooperate synergistically in a variety of biological functions, including protein folding. The properties of DnaJ and its homologs have been reviewed previously (23, 34, 131, 260).

Little is known about the structural features of DnaJ that are involved in its interaction with DnaK and unfolded proteins. Analysis of DnaJ fragments showed that both the NH_2-terminal J domain and the adjacent glycine/phenylalanine-rich region are required for interactions with DnaK (117) and to stimulate the ATPase activity of DnaK (220). The G/F motif of DnaJ is also involved in modulating the substrate binding activity of DnaK (246). However, only complete DnaJ is functional with DnaK and GrpE in refolding denatured firefly luciferase. Binding experiments and cross-linking studies indicate that the zinc fingerlike domain is required for DnaJ to bind to nonnative proteins (220).

Nuclear magnetic resonance spectroscopy has been used to determine the three-dimensional structure of the J domain in DnaJ from *E. coli* and humans (101, 181, 222). The structure is dominated by two long helices, with a hydrophobic core of highly conserved side chains. The residues believed responsible for the specificity of the interaction between DnaJ and its homologs with their corresponding HSP70 partners comprise a conserved His-Pro-Asp sequence that extends out from the core of the structure (171, 181). A peptide containing this sequence inhibited the Ydj1 stimulation of HSP70 ATPase activity but did not prevent binding of nonnative substrate proteins, indicating that DnaJ interacts with HSP70 at a site distinct from the peptide binding site (238). The adjacent Gly/Phe-rich domain in DnaJ is disordered and flexible in solution (222).

The major effect of DnaJ on the functional cycle of DnaK is the significant stimulation of the ATPase rate-limiting step, γ-phosphate cleavage, leading to stabilization of DnaK-ADP-substrate protein complexes (146). Both prokaryotic and eukaryotic forms of HSP40 interact with HSP70 in the presence of ATP to suppress protein aggregation (32). It has been proposed that HSP40 is required for the efficient binding of substrate protein to HSP70 through the stimulation of its ATPase activity (see sect. v*A*) (147).

It has been suggested DnaJ acts directly as a molecular chaperone in that it binds to certain denatured substrate proteins such as firefly luciferase (130, 220, 221), and even some specific folded proteins such as the σ^{32}-heat shock transcription factor or the λP DNA replication protein, but not "normal" native proteins (41, 262). However, DnaJ binds to σ^{32} at a different site than that to which DnaK binds. The yeast DnaJ homolog Ydj1 was found to bind to denatured rhodanese but not unfolded reduced carboxymethylated α-lactalbumin (32, 35). As discussed in section iv, DnaJ or HSP40 has been proposed to bind to nascent polypeptides to prevent their premature folding and to target HSP70 to them (70). However, unambiguous data to support this role are scant. In yeast, DnaJ and its homologs are required not only for protein folding but also for selective ubiquitin-dependent degradation of abnormally folded proteins (132).

Significant specificity in the interactions between members of the HSP70, DnaJ, and GrpE families has been observed (40). As noted, the interaction between a given HSP70 and its interacting DnaJ is determined by the J domain (198). Recently, evidence for interactions between DnaJ homologs and HSP90 have been reported (118). It has also been suggested that DnaJ possesses an active dithiol/disulfide group and may catalyze protein disulfide formation, reduction, and isomerization (38).

C. HSP60 Family

Under the rubric of the HSP60 or chaperonin family, we consider both the GroEL and TCP-1 ring complex families. Unfortunately, different research groups have used different names for the TCP-1 ring complex, e.g., TRiC (for TCP-1 ring complex) and CCT (for chaperonin containing TCP-1). Other members include the Rubisco subunit binding protein and thermophilic factor 55 from archaea. GroEL and its homologs are found in prokaryotes, chloroplasts, and mitochondria, whereas TCP-1 and its homologs are found in the eukaryotic cytosol. Many of the HSP60 chaperones are also known as chaperonins (cpn60) and are ring-shaped oligomeric protein complexes with a large central cavity in which nonnative proteins can bind. In bacteria, at least, HSP60 require a cochaperonin, GroES (cpn10), for full function. The term *chaperonin* was originally coined by Ellis (48) to refer to non-heat-induced HSP60.

GroEL is probably the most studied of all molecular chaperones; in combination with its cochaperonin GroES and ATP, it facilitates protein folding, not only by preventing aggregation but also by simultaneously allowing partially folded intermediates to fold in an environment conducive to stabilizing the native state. It has been suggested that GroEL may also function by unfolding misfolded states so as to allow their productive refolding (268, 269). Members of the HSP60 family are also involved in the assembly of large multiprotein complexes such as Rubisco (27, 243). The availability of a high-resolution crystallographic structure, in conjunction with mutagenesis studies, has helped in the elucidation of the details of the reaction cycle (see sect. v*B*). However, there are still many points of controversy, reflecting the complexity of the mechanism of this large chaperone. Recent reviews include References 53, 105, 108, 144.

The structure of the *E. coli* chaperonin GroEL has been solved by X-ray crystallography (9, 12, 263) and electron microscopy (196) and consists of 14 identical subunits in two stacked heptameric rings, each containing a central cavity. Substantial structural information about GroEL, GroES, and related chaperonins is available from the chaperonin web home page: http://bioc09.uthscsa.edu/~seale/Chap/struc.html. Each subunit consists of three domains: the equatorial, the intermediate, and the apical. The latter, forming the mouth of the central cavity, undergoes major conformational changes on binding of ATP and the cochaperonin GroES, which lead to substantial changes in the hydrophobic nature of the cavity (263) (Fig. 2). In particular, the relatively hydrophobic cavity lining to which the unfolded substrate protein binds before GroES binding becomes much more polar, coincident with a substantial increase in the size of the cavity. The hydrophobic polypeptide-binding site on the cavity-lining surface of the apical domain was identified with the help of various mutants (54). These same residues are also essential for binding of the cochaperonin GroES, which is required for productive polypeptide release.

The identity of amino acid residues at the nucleotide-binding sites of GroEL/GroES was determined by photoaffinity labeling with 2-azido-ATP (13). The labeled site is located at the GroEL/GroEL subunit interface, and labeling of the cochaperonin GroES occurred through a conserved proline. The 2.4-Å crystal structure of the bacterial chaperonin GroEL complexed with adenosine 5′-*O*-(3-thiotriphosphate) bound to each subunit shows that ATP binds in a pocket with a unique nucleotide-binding

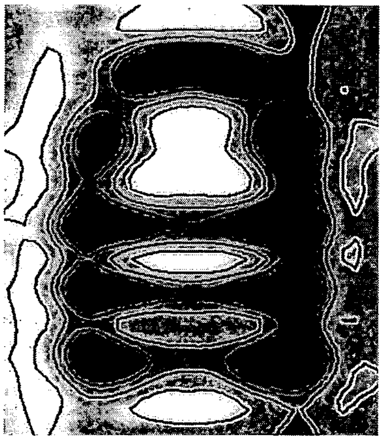


FIG. 2. Structure of GroEL/GroES based on electron diffraction. A cross section through a stacked GroES/GroEL/GroEL unit is shown. Main features starting at top are GroES; enlarged cavity; boundary between stacked GroEL units, flanked by equatorial domains; and smaller cavity, with apical domain pointing inward. Major change in cavity size upon binding GroES is clearly visible. [From Chen et al. (26).]

motif, whose primary sequence is highly conserved among chaperonins (9).

The 45-Å-diameter cavity in GroEL is large enough to accommodate proteins of 40–50 kDa and, as noted, is larger when capped by the GroES heptamer. Conformational changes are observed on binding nucleotides or GroES. The hydrophobic nature of the central cavity in GroEL (in the absence of the cochaperonin) presumably accounts for the lack of affinity for native proteins. GroES enhances the cooperativity of ATP binding and hydrolysis by GroEL and is necessary for the release and folding of many GroEL substrates. The crystallographic structure of GroES is known to high resolution (110). GroES has a highly mobile and accessible polypeptide loop whose mobility and accessibility are lost upon formation of the GroES/GroEL complex (128, 129).

The TCP-1 is a heteroligomeric 970-kDa complex containing several structurally related subunits of 52–65 kDa found in the eukaryotic cytosol. These are assembled into a ring complex that resembles the GroEL double ring (69, 121, 187). In vitro, the TCP-1 ring complex appears to function independently of a small cochaperonin protein such as GroES. Thus far, TCP-1 complexes have been shown to be involved in the folding of very few proteins in the eukaryotic cytosol.

The major difference between TCP-1 complexes and GroEL is the heteroligomeric nature of the TCP-1 ring complex; at least eight subunit species that are encoded by unique genes are known (216). The genes are calculated to have diverged around the starting point of the eukaryotic lineage and share ~30% amino acid identity. It has been proposed that this complexity may have evolved to cope with the folding and assembly of complex proteins in eukaryotic cells (122). Although each TCP-1 subunit is highly diverged from each other, individually they are quite homologous, suggesting that each subunit has a specific, independent function (30a,121).

D. HSP70 Family

The HSP70 are a family of molecular chaperones that are involved in protein folding and several other cellular functions and that exhibit weak ATPase activity. The HSP70 chaperones are composed of two major functional domains. The NH_2-terminal, highly conserved ATPase domain binds ADP and ATP very tightly (in the presence of Mg^{2+} and K^+) and hydrolyzes ATP, whereas the COOH-terminal domain is required for polypeptide binding. Cooperation of both domains is needed for protein folding. Several recent reviews summarize the role of HSP70 molecular chaperones in protein folding (58, 75, 83, 90, 100, 144). Many of the functions of the *E. coli* HSP70, DnaK, require two cofactors, DnaJ (see sect. IIIB) and GrpE (see sect. IIIJ). The majority of in vitro studies on HSP70 have been with DnaK.

The HSP70 family is very large, with most organisms having multiple members; most eukaryotes have at least a dozen or more different HSP70, found in a variety of cellular compartments. Some of the better known mammalian members are HSC70 (or HSP73), the constitutive cytosolic member; HSP70 (or HSP72), the stress-induced cytosolic form; BiP (or Grp78), the ER form; and mHSP70 (or mito-HSP70, or Grp75), the mitochondrial form. In yeast the homologs of HSC70 and BiP are known as Ssa1–4 and Kar2. In *E. coli*, the major form of HSP70 is DnaK. Here we will use the term *HSP70* to refer to any member of the family.

The crystallographic structures of the bovine HSC70 ATPase domain, the DnaK peptide-binding domain complexed with a peptide substrate, and most recently the human HSP70 ATPase domain have been determined (63, 214, 272). The ATPase domain, which is structurally similar to actin and hexokinase, consists of four smaller domains forming two lobes with a deep cleft within which the MgATP and MgADP bind. The structure of the peptide-binding domain consists of a β-sandwich subdomain followed by α-helical segments. The peptide is bound to DnaK in an extended conformation through a channel defined by loops from the β-sandwich. An α-helical domain (the flap or latch) is believed to stabilize the complex but does not contact the peptide directly. Only five residues of the substrate protein make significant contacts with HSP70, explaining the previously observed specificity for short, hydrophobic peptides, with a strong preference for hydrophobic residues such as Leu in the central region and a strong unfavorable interaction with negatively charged residues (7, 64, 80, 225). A model in which the flap over the substrate binding pocket could be in either an open conformation (to allow entry and egress of substrates) or closed conformation (to form a stable complex) has been suggested to account for the high- and low-affinity states of HSP70 (272).

Recently, the substrate specificity of DnaK has been mapped out in detail by screening an immobilized peptide library (192), following up on earlier peptide-scanning experiments (7). DnaK binding sites in protein sequences occurred statistically every 36 residues. In the folded proteins, these sites are mostly buried, and the majority are found in β-sheets. The binding motif consists of a hydrophobic core of four to five residues enriched particularly in Leu, but also in Ile, Val, Phe, and Tyr, and two flanking regions enriched in basic residues. Acidic residues are excluded from the core and disfavored in flanking regions. On the basis of these data, an algorithm was established that predicts DnaK binding sites in protein sequences with high accuracy (192).

The HSP70 preferentially bind unfolded or partially folded proteins and do not bind normal native proteins (although there are a few specific interactions with proteins in their native states, such as clathrin and σ^{32}). It is likely that only some newly synthesized proteins require the assistance of chaperones. In coimmunoprecipitation studies with anti-HSP70 antibodies and pulse-chase labeling, it was observed that smaller proteins were disproportionately absent, suggesting that they may fold more rapidly, either with or without the assistance of HSP70 (5).

In fact, there is some evidence to support the notion that HSP70 may not interact with short-lived partially folded intermediates (218). The HSP70 inhibits the refolding of the mitochondrial isozyme of aspartate aminotransferase (AAT), but not the cytosolic homolog. This has been attributed to HSP70 binding to a long-lived early folding intermediate in the folding of mitochondrial AAT, for which the analogous cytosolic isozyme intermediate is shorter lived and rapidly transforms to a more nativelike species that does not bind to HSP70 (3). Because there will always be a kinetic competition between spontaneous folding and chaperone binding, intermediates with shorter lifetimes than that required for binding to HSP70 would not form a complex with the chaperone (Fig. 3).

The rapid binding kinetics for substrate proteins to DnaK-ATP (199) suggest that ATP-bound DnaK is the primary form initiating interaction with substrates for chaperone activity. The resulting DnaK-ATP-substrate complexes, however, are also characterized by rapid dissociation of bound substrate but can be stabilized by hydrolysis of the ATP (stimulated to a small extent by the substrate itself, or to a large extent by DnaJ; Ref. 146). The ATP-induced protein-HSP70 complex dissociation results from a conformational change induced in HSP70 by ATP binding. This conformational change decreases the affinity of HSP70 for nonnative substrate proteins and leads to their dissociation (166). Because the binding of ATP occurs in the NH_2-terminal domain and peptide binding is in the COOH-terminal domain, it is clear that strong coupling between the two functional domains must exist.

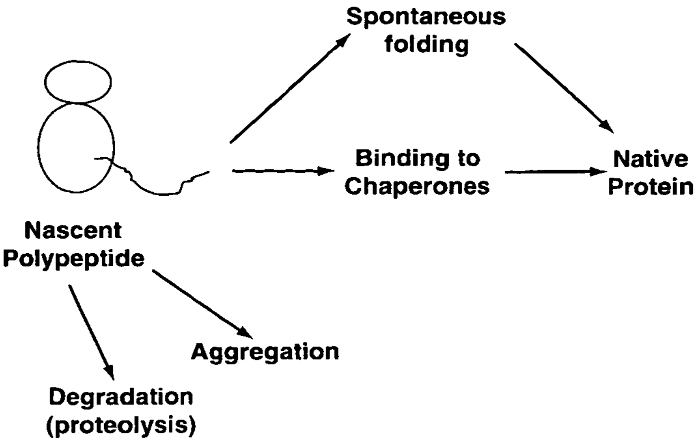


FIG. 3. A newly synthesized protein, whether still associated with ribosome or released, faces several competing pathways. It is likely that in the absence of chaperones, aggregation or other forms of misfolding would be the major pathway for many proteins.

Under appropriate conditions, DnaK undergoes autophosphorylation (137). It is not yet clear if this is of physiological significance. At the moment, there is no good evidence that it is. However, GrpE and synthetic peptides have been observed to inhibit the phosphorylation (the effectiveness of a given peptide correlated with its affinity for DnaK), whereas DnaJ had no effect on the reaction (169). Human HSP70 is phosphorylated in vitro in the presence of divalent ions, with calcium being the most effective. Two calcium ions were found in the human ATPase domain structure, and calcium binding may facilitate phosphorylation (214).

Various techniques have shown that HSP70 adopts at least three significantly different conformations, one in the absence of nucleotide, one with ADP bound, and one with ATP bound. Binding of nucleotides or polypeptides alters the conformations of both the nucleotide- and polypeptide-binding domains, further indication that the conformations of these two domains are highly coupled (71).

Recently, a new pair of DnaK/DnaJ-like chaperones has been discovered in *E. coli* (242). Sequence differences between HSC66 and HSC20 compared with other HSP70/HSP40 members suggest that these chaperones may have different peptide binding specificity and be subject to different regulatory mechanisms. In particular, the high level of constitutive expression and lack of significant response to temperature changes suggest that HSC66 and HSC20 may play an important role in the folding of certain newly synthesized proteins under normal cellular conditions.

Details of the mechanism by which HSP70 interact with newly synthesized and nonnative proteins are given in sections IV and V*A*.

E. HSP90 Family

Members of the HSP90 family are highly conserved, essential proteins found in all organisms from bacteria to humans. Examples include the cytosolic form in eukaryotes, HSP90, the ER form, Grp94, and the *E. coli* homolog HtpG. Mammalian HSP90 exist as dimers. Although there are a number of similarities between the activities of HSP90 and HSP70, the former has several identified specific interactions, for example, with cytoskeleton elements, signal transduction proteins (including steroid hormone receptors), and protein kinases (such as the mitogen-activated protein kinase system). HSP90 is frequently found in complexes with other chaperones. In vitro, HSP90 exhibits chaperone activity with diverse proteins, suggesting a general function. The properties of HSP90 have been reviewed (10, 11, 20, 113, 175, 265).

Recently, the crystal structure of the NH_2-terminal domain of the yeast HSP90 was solved to reveal a dimeric structure based on a highly twisted 16-stranded β-sheet. The opposing faces of the β-sheet in the dimer define a potential peptide-binding cleft, suggesting that the N domain may serve as a molecular "clamp" in the binding of ligand proteins to HSP90 (179).

There has been a long-standing controversy as to whether HSP90 binds or hydrolyzes ATP. The crystal structures of complexes between the NH_2-terminal domain of the yeast HSP90 with ADP/ATP unambiguously show a specific adenine nucleotide binding site, homologous to the ATP-binding site of DNA gyrase B. This site is the same as that identified for binding the antitumor agent geldanamycin, suggesting that geldanamycin acts by blocking the binding of nucleotides to HSP90 and not the binding of incompletely folded substrate proteins as previously suggested. These results strongly suggest the direct involvement of ATP in the function of HSP90 (82, 178).

Even though HSP90 is one of the most abundant chaperones in the cell, its in vivo functions are poorly understood, and little is currently known about its role in chaperoning the folding of newly synthesized proteins, although there are hints that it does not function alone but is associated with several other cofactors. For example, HSP90 performs at least part of its function in a complex with members of the prolyl isomerase family, FKBP52 and p23 (10), and the steroid receptor complex consists of HSP90, HSP70, p48, the cyclophilin Cyp-40, and the associated proteins p23 and p60 (45). Although neither Cyp-40 nor p23 can refold unfolded substrates, in in vitro folding experiments they interact with nonnative proteins and maintain a folding-competent intermediate (67).

A temperature-sensitive mutant of HSP90 in yeast, which rapidly and completely loses activity on shift to high temperatures, has been used to examine the functions of HSP90 in vivo. The results suggested that HSP90 is not required for the de novo folding of most proteins but is required for a specific subset of proteins that have greater difficulty reaching their native conformations (153). In vitro, in the absence of nucleotide, HSP90 can maintain nonnative substrate in a "folding-competent" state that refolds upon addition of HSP70, DnaJ homolog, and nucleotide (66).

F. HSP100 Family

The heat-inducible members of the HSP100 (or Clp) family of proteins have a number of very intriguing properties and share a common function in helping organisms to survive extreme stress (78). They perform a diverse set of functions, including proteolysis. They are highly conserved, present in all organisms, and contain ATP and polypeptide binding sites. Both HSP104 and ClpA form six-membered ring complexes; the diameter of the interior of the rings is much smaller than in GroEL, making it unlikely that the HSP100 function analogously to HSP60. The basic mechanisms by which these chaperones function are not understood. There is some suggestion that HSP104 may act in concert with HSP70 and DnaJ homologs to increase the yields of renatured protein (78). It should be noted that no human analogs of HSP104 have been found.

Unlike HSP60 and HSP70, which are unable to resolubilize aggregated proteins in vitro (with the exception of RNA polymerase), HSP104 has been observed to solubilize thermally aggregated proteins both in vivo and in vitro (170). Interestingly, ClpA can substitute for the ATP-dependent chaperone function of DnaK and DnaJ in the in vitro activation of the plasmid P1 RepA replication initiator protein (257). Another unusual feature of HSP104 is its role in triggering a prionlike disorder in yeast, involving the extrachromosomal elements PSI+ and URE3 (37).

G. Calnexin and Calreticulin

Calnexin is a transmembrane molecular chaperone that resides in the ER. Calreticulin, which has sequence homology with calnexin, is a soluble ER chaperone. Both proteins are involved in the folding and assembly of nascent proteins in the ER in a calcium-dependent manner and play an important role in glycoprotein maturation and quality control in the ER (6, 93, 119, 259).

Most proteins that enter the ER are cotranslationally modified by the addition of a complex carbohydrate structure that undergoes subsequent modification by selective removal of individual hexose residues (219). Both calreticulin and calnexin transiently interact with many newly synthesized proteins in the ER, with some overlap between those proteins that bind to calreticulin and those that bind to calnexin. The specificity of the interaction is determined by the nature of the oligosaccharide and requires the trimming of glucose residues from the asparagine-linked core glycans by glucosidases. Calnexin transiently interacts with newly synthesized glycoproteins, specifically recognizing a monoglucosylated intermediate (162). Its major role appears to be to monitor glycoprotein folding and prevent incompletely folded proteins from leaving the ER. As proposed by Helenius and co-workers (92), carbohydrate processing and folding occur simultaneously; calnexin recognizes and binds the monoglucosylated glycoprotein intermediate of the nascent chain. After the remaining glucose is removed, the glycoprotein is released from calnexin; if it is incompletely folded, it is reglycosylated and rebinds to calnexin. If it is folded it no longer binds to calnexin. Calreticulin is also specific for monoglucosylated glycans (172). The interactions of calreticulin and calnexin with denatured proteins are highly dependent on divalent metal ions or polyamines (261). Calnexin facilitates the folding and assembly of class I histocompatibility molecules and prevents formation of aggregates, showing that it functions as a molecular chaperone (241). Protein folding in the ER is also discussed in section IV*A*.

H. Protein Disulfide Isomerase

Protein disulfide isomerase (PDI) is a critical cofactor in the folding of many proteins that are found in the ER (65, 77, 143, 176). Many secreted proteins have multiple disulfide bonds, presenting potential problems for correct disulfide pairing during folding. In vitro studies of the refolding of reduced proteins show that disulfide bond formation occurs rapidly and is followed much more slowly by thiol-disulfide rearrangement leading to the correct disulfide pairings. Thus catalysis of oxidative folding is necessary in vivo to rapidly generate the correct disulfide bonds in newly synthesized proteins. In the eukaryotic ER, PDI fulfills this function. Its concentration can reach close to millimolar levels. The properties of PDI have recently been reviewed (245). In addition to strong affinity for unfolded proteins and peptides, it binds many relatively hydrophobic molecules such as steroid and thyroid hormones. Hence, it is not surprising that PDI has been reported to have chaperone-like activity at high concentrations (such as inhibition of aggregation) distinct from its disulfide bond interactions (22, 180, 209).

Protein disulfide isomerase has two catalytic sites situated in two domains homologous to thioredoxin, one near the NH_2 terminus and the other near the COOH terminus. The thioredoxin domains, by themselves, can catalyze disulfide formation, but they are unable to catalyze disulfide isomerizations (36).

I. Peptidyl Prolyl Isomerase/Trigger Factor

Under in vitro (and presumably in vivo) conditions, proline *cis-trans* isomerization may become rate limiting in the folding of proteins; in many cases, the presence of peptidyl prolyl isomerase (PPI) will enhance the rate of folding. Peptidyl prolyl isomerases are ubiquitous enzymes found in virtually all organisms and subcellular compartments. Three unrelated families are known: the cyclophilins, the FK506-binding proteins (FKBP), and the parvulins (200). The former two families are also known as immunophilins. The trigger factor is a PPI with somewhat similar activity, and weak homology, to FKBP. Trigger factor is an abundant cytosolic protein originally identified by its ability to maintain the precursor of a secretory protein in a translocation-competent form (31).

Structural studies of the *E. coli* trigger factor reveal a modular structure, composed of three stably folded domains, of which the catalytic one is homologous to FKBP (270). Trigger factor binds partially folded intermediates tightly. Although the isolated catalytic domain of the trigger factor retains full prolyl isomerase activity toward short peptides, its activity toward protein substrates is dramatically reduced, indicating that the polypeptide binding site extends beyond the FKBP domain (201).

Trigger factor has several chaperone-like functions: it binds to nascent cytosolic and secretory polypeptide chains, and it catalyzes protein folding in vitro (98). Trigger factor interacts with GroEL in vivo and promotes its binding to at least some polypeptides; GroEL-trigger factor complexes show much greater affinity for partially folded intermediates than GroEL alone (116). On the basis of studies showing that trigger factor was cross-linked to all tested nascent chains derived from both secreted and cytosolic proteins, it appears that trigger factor may act as a general molecular chaperone in protein synthesis (99, 240).

J. HSP70 Cochaperones

In addition to DnaJ and GrpE, which function as cochaperones with DnaK, and have been known for several years, other protein cofactors that interact with HSP70 have been discovered recently. These include Hip (HSC70-interacting protein), BAG-1, and auxilin. The existence of these cofactors illustrates the complexity of the HSP70 chaperone machinery in cells.

GrpE is a key component of the HSP70 chaperone system for protein folding in bacteria and mitochondria. GrpE acts as a nucleotide exchange factor to control the ATPase activity of DnaK in its reaction cycle, although the details of its mechanism remain unclear. GrpE has high affinity for monomeric native DnaK, as well as the isolated ATPase domain (185, 202). GrpE has no affinity for ATP or ADP, nor the oligomeric states of DnaK. The nucleotide exchange properties of GrpE are a consequence of the binding of GrpE to DnaK, leading to a conformational change involving the opening of the nucleotide cleft on DnaK, resulting in a low-affinity state for nucleotides. Recently, the crystal structure of GrpE bound to the ATPase domain of the molecular chaperone DnaK has been determined (89). A dimer of GrpE binds asymmetrically to a single molecule of DnaK. The structure of the nucleotide-free ATPase domain complexed with GrpE closely resembles that of the nucleotide-bound mammalian HSP70 homolog, except for an outward rotation of one of the subdomains of the protein. Two long α-helices extend away from the GrpE dimer and suggest an additional role for GrpE in peptide release from DnaK. The functional aspects of GrpE are given in section *vA*.

Hip is a novel tetrameric cochaperone involved in the regulation of eukaryotic HSC70, distinct from that of bacterial HSP70. It appears to play a role in forming stable HSP70 complexes with substrate proteins. One Hip oligomer binds the ATPase domains of at least two HSC70 molecules, dependent on activation of the HSC70 ATPase by HSP40. Although hydrolysis remains the rate-limiting step in the ATPase cycle, Hip stabilizes the ADP state of HSC70 that has a high affinity for substrate protein. Hip also appears to be a chaperone in its own right, in that it binds to some unfolded proteins (15, 104).

BAG-1 is a recently discovered regulator of HSP70 (216, 271). BAG-1 is an antiapoptotic protein and also interacts with several steroid hormone receptors that require the molecular chaperones HSP70 and HSP90 for activation. The action of BAG-1 is similar to that of GrpE in bacterial cells, in that it binds to the ATPase domain of HSP70 and, in cooperation with HSP40, stimulates the rate of ATP hydrolysis by increasing the rate of release of ADP from HSP70 (103). BAG-1 can be coimmunoprecipitated with HSP70 from cell lysates (223). BAG-1 inhibited the HSP70-mediated in vitro refolding of an unfolded protein substrate. The binding of BAG-1 to one of its known cellular targets, Bcl-2, in cell lysates was found to be dependent on ATP, consistent with the possible involvement of HSP70 in complex formation. The identification of HSP70 as a partner protein for BAG-1 may explain the diverse interactions observed between BAG-1 and several other proteins, including steroid hormone receptors and certain tyrosine kinase growth factor receptors.

Auxilin is a 100-kDa cofactor involved in the HSP70-mediated uncoating of clathrin-coated vesicles (239). Clathrin-coated vesicles transport selected integral membrane proteins from the cell surface and the *trans*-Golgi network to the endosomal system. Before fusing with their target, the vesicles must be stripped of their coats. Auxilin binds with high affinity to assembled clathrin lattices and, in the presence of ATP, recruits HSP70. The presence of a J domain at its COOH terminus indicates that auxilin is a member of the DnaJ family: deletion of the J domain results in the loss of cofactor activity.

A 16-kDa cytosolic protein, called p16, which copurifies with HSC70 from fish liver, has been identified as a member of the Nm23/nucleoside diphosphate kinase family (135). p16 may modulate HSC70 function by maintaining HSC70 in a monomeric state and by dissociating unfolded proteins from HSC70 either through protein-protein interactions or by supplying ATP indirectly through phosphate transfer.

Hop is a recently discovered 60-kDa protein that can form a physical link between HSP70 and HSP90, thus modulating their activities (114). Hop is involved in the refolding of denatured protein in rabbit reticulocyte lysate and stimulates the refolding by HSP70 and Ydj-1 in a purified refolding system. Optimal refolding was observed in the presence of both Hop and HSP90. Hop preferentially formed a complex with ADP-bound HSP70 and also appears to bind to the ADP-bound form of HSP90.

K. Specialized Chaperones

Some molecular chaperones may be highly specific in that they interact with only one, or a very limited number, of target proteins; examples are PapD (127), which is involved in the assembly of bacterial pili, and HSP47, which is involved in the folding and processing of procollagen in the ER. There are many large and complex protein machines in cells: in some of these cases, specific molecular chaperones are involved in their assembly (206). Some of the best-studied systems are bacteriophage capsids and bacterial pili and flagella.

The 47-kDa HSP (HSP47) is an ER-resident chaperone found in collagen-producing cells, where it interacts with procollagen. It has been proposed that it functions as a chaperone regulating procollagen chain folding and/or assembly, but the mechanism is not well understood (152). It is likely that its main function is to prevent aggregation and misfolding of newly synthesized procollagen chains until the correct COOH-terminal associations have been made to yield the collagen triple helix. When HSP47-procollagen complexes reach the *cis*-Golgi network, the chaperone rapidly dissociates. The major interaction site on procollagen has been shown to be the pro-alpha 1 N-propeptide (109).

Receptor-associated protein (RAP) is another example of a specialized molecular chaperone, in this case for the low-density lipoprotein receptor-related protein (LRP), a large receptor that binds multiple ligands. The major role of RAP is to facilitate correct folding of LRP and to prevent the premature interaction of ligands with LRP (161).

The production of native α/β-tubulin heterodimer depends on the action of cytosolic chaperonin and at least five protein cofactors. These reactions do not depend on ATP hydrolysis (230, 231). The β-tubulin monomer release factor, p14, which catalyzes the release of β-tubulin monomers from intermediate complexes, has recently been shown to be a member of the DnaJ family (141).

There are also a number of chaperones involved in protein export, such as SecB from *E. coli* (88). SecB has two functions: it maintains precursors of some exported proteins in a conformation compatible with export, by preventing them from aggregating or from folding to their native state in the cytoplasm, and it delivers both nascent and completed precursors to SecA, one of the components of the export apparatus associated with the plasma membrane. Only those polypeptides that fold slowly interact significantly with SecB, even though it is able to bind a wide variety of nonnative proteins. Complexes between SecB and substrate proteins are in rapid equilibrium with the free states (236). Thus, unlike the HSP70 and HSP60, in which hydrolysis of ATP is coupled to the binding and release of substrate proteins, SecB does not form stable complexes with substrate proteins. This may reflect the fact that SecB does not mediate protein folding but is specialized for the protein export pathway.

IV. INTERACTIONS OF NASCENT CHAINS WITH CHAPERONES

Fundamental questions in protein biogenesis include at which stage the nascent protein first interacts with molecular chaperones, the identity of the chaperones, and the role of the chaperones in facilitating protein folding. Do they just prevent aggregation and misfolding, or do they play a more active role in the actual folding process? The involvement of chaperones in both co- and posttranslational folding is now clear.

Considerable controversy continues regarding which chaperones are involved in interactions with nascent polypeptide chains. The initial evidence for chaperoning came from studies on the assembly of immunoglobulin light and heavy chains and the involvement of the protein now known as BiP (an HSP70) (85, 151). In a study with major implications for in vivo protein folding and assembly, Welch and co-workers (5) demonstrated that cytosolic forms of HSP70 bind cotranslationally to nascent polypeptide chains and to newly synthesized proteins in the normal (unstressed) cell in an ATP-dependent manner. The association of cytosolic HSP70 with nascent polypeptide in translating ribosomes has subsequently been confirmed in a number of organisms (154).

There have been several reports that a high-molecular-weight complex of proteins including various chaperones is associated with nascent (or unfolded) polypeptide chains during chain elongation in vitro and in vivo. Early evidence was observed in the renaturation of firefly luciferase in cell-free translation systems (70, 97, 205). Chaperone-stabilized luciferase was associated with high-molecular-weight complexes overlapping the distributions of HSP70, HSP90, and the chaperonin TRiC on gel filtration columns (160). Molecular chaperones that have been implicated include HSP70 (5, 86); HSP70 and HSP40 (123, 154); HSP70, HSP40, and the TCP-1 ring complex chaperonin (70); and HSP70 and HSP90 (46, 205).

In a clever new approach, an antibody to puromycin was used to identify a population of truncated nascent polypeptides that were then probed by immunoprecipitation and chemical cross-linking with several antibodies that recognize the cytosolic chaperones HSP70, CCT (TRiC), HSP40, p48 (Hip), and HSP90, as a means of identifying chaperones bound to the nascent chains (46). The results showed that HSP70 is the predominant chaperone bound to nascent polypeptides. The interaction between HSP70 and nascent polypeptides is apparently dynamic under physiological conditions but can be stabilized by depletion of ATP or by chemical cross-linking. Interestingly, the cytosolic chaperonin CCT (TRiC) was found to bind primarily to full-length, newly synthesized actin and tubulin. Other studies have also implicated the TCP-1 ring complex in the synthesis and assembly of tubulin and actin (69, 215, 264).

This investigation also demonstrated that nascent polypeptides have a strong propensity to bind to many proteins nonspecifically in cell lysates. It is likely that this nonspecific binding is responsible for the reports of additional components in contact with nascent polypeptides (46).

Several studies provide support for cotranslational interactions of molecular chaperones with nascent polypeptide chains, especially HSP70 and perhaps HSP40. The interaction of DnaJ with nascent ribosome-bound polypeptide chains as short as 55 residues was reported using firefly luciferase and chloramphenicol acetyltransferase in cross-linking experiments (97). These investigations showed that both folding and subsequent mitochondrial translocation required DnaK, DnaJ, and GrpE and led to the proposal that DnaJ protects nascent polypeptide chains from aggregation and, in cooperation with HSP70, controls their productive folding once a complete polypeptide or a polypeptide domain has been synthesized. Both HSP70 and HSP40 were shown to be associated with nascent polypeptide chains in translating ribosomes, whereas GroEL, although transiently associated with newly synthesized proteins, was absent from the ribosomes, suggesting that HSP70 and HSP40 play an early role in protein folding, whereas GroEL acts at a later stage (70, 72, 173).

Investigations using fluorescent-labeled rhodanese (by the cotranslational incorporation of a coumarin derivative at the NH_2 terminus of the nascent protein) demonstrated the accumulation of full-length but enzymatically inactive polypeptides on the ribosomes. These polypeptides could be activated and released by subsequent incubation with the chaperones DnaJ, DnaK, GrpE, GroEL, GroES, and ATP and release factor. Changes in fluorescence indicated that DnaJ bound to the nascent protein and appeared to be essential for folding of ribosome-bound rhodanese into the native conformation (87, 124, 126).

Further support that folding of nascent proteins can take place on the ribosome comes from studies on partially folded intermediate states of bacteriophage P22 tailspike protein and the β-subunit of tryptophan synthase, which can be detected while still bound to ribosomes using monoclonal antibodies to the intermediates. The rapid appearance of the intermediates suggests that the nascent chains start folding during their elongation on the ribosomes. The newly synthesized incomplete chains were shown to interact with DnaK but not GroEL while still bound to the ribosome (235).

There has been considerable discussion as to whether GroEL interacts with newly synthesized proteins in a cotranslational or posttranslational manner. As noted above, several studies indicated that DnaK and DnaJ are involved at an early stage in the folding of newly synthesized protein and that GroEL acts at a later stage (72). In a recent investigation in which rhodanese was synthesized in both in bacterial and wheat germ translation extracts, only posttranslational stable complexes with GroEL were found (184). Further evidence consistent with the HSP70 chaperone machinery interacting with newly synthesized proteins before GroEL (or concurrently) comes from investigations on the synthesis of chloramphenicol acetyltransferase in a system genetically depleted of DnaK and DnaJ. Most of the chloramphenicol acetyltransferase failed to assemble into active trimers and accumulated either in a complex with GroEL or as inactive monomer. The addition of DnaK and DnaJ to the system before the start of protein synthesis led to increased formation of native chloramphenicol acetyltransferase (244). Another investigation supporting a place for GroEL in the later stages of the folding of newly synthesized proteins made use of temperature-sensitive lethal mutations in the GroEL gene. After a shift to a nonpermissive temperature, the rate of general translation in the mutant cells was reduced, but a specific group of cytoplasmic proteins failed to fold to their native states (107). The much more limited specificity demonstrated for TCP-1 chaperonins, compared with GroEL, suggests significantly different roles for these two classes of chaperonins in the biosynthesis of proteins. It is likely that the functional differences reflect underlying structural differences.

Bukau and co-workers (16) have recently suggested that during the folding of newly synthesized proteins, DnaK and GroEL do not act in sequence, but rather the two chaperone systems form a "lateral network of cooperating proteins." There are data to support both this and the sequential models, so the question remains unresolved at present. Another source of controversy relates to the question of how many newly synthesized proteins require the assistance of HSP60 (chaperonins) in folding. Lorimer has calculated that for *E. coli* there is only sufficient GroEL to assist ~5% of newly translated proteins under normal conditions (142). It is therefore likely that most newly synthesized proteins in *E. coli* fold without the assistance of GroEL, and this implies that most proteins fold fast enough that sequestration on DnaK to minimize the concentration of nonchaperone-bound protein suffices to prevent aggregation.

Thus, whereas the overall outline of the process of chaperone-mediated folding of newly synthesized proteins is clear, the details are as yet incompletely resolved. A nascent polypeptide will interact with HSP70 and possibly other chaperones (probably HSP40) as it emerges from the ribosome. The lifetime of HSP70 complexes with substrate proteins under in vivo conditions is not well established but is likely to be comparable to the time for folding of many newly synthesized proteins. Dissociation of the newly synthesized chain from HSP70 after release of the nascent chain from the ribosome sets up a kinetic competition between rebinding to HSP70, binding to HSP60, spontaneous folding, aggregation, or possible even proteolysis (Fig. 3).

It has also been reported that there may be significant differences between folding in prokaryotes and eukaryotes (155). In a eukaryotic translation system, two-domain engineered polypeptides were observed to fold by sequential and cotranslational folding of their domains. However, in *E. coli*, folding of the same proteins was found to be posttranslational and to lead to intramolecular misfolding of the concurrently folding domains (155). In addition, differences between the in vitro and in vivo nature of the interactions of chaperones with actin during refolding from denaturant have been reported (68).

A. Folding in the Endoplasmic Reticulum

The ER is a key compartment in cells that are specialized for protein export and contains many chaperones that are essential for the production of functional proteins for export (250). Folding begins with the insertion of a preprotein into the lumen of the ER and can occur either posttranslationally, in which case the preprotein is completely synthesized on cytosolic ribosomes before being translocated, or cotranslationally, in which case membrane-associated ribosomes direct the nascent polypeptide chain into the ER concomitant with polypeptide elongation (14). The ER has excellent quality control mechanisms (involving chaperones) that recognize and selectively retain misfolded proteins, which are then either degraded or refolded (30, 149). The concentration of the ER HSP70, BiP, is increased by elevated levels of misfolded proteins in the ER. How the levels of misfolded molecules are monitored and how this information is used to regulate the synthesis of BiP are still poorly understood. Likewise, the mechanisms by which oxidizing potential of the ER environment is regulated, and the misfolded proteins are degraded, are also unknown.

Although some of the major chaperones involved in protein folding in the ER are well studied, e.g., BiP and PDI, it is apparent that more have yet to be characterized. For example, several calcium-dependent putative chaperones have recently been identified using affinity chromatography with denatured-protein columns and elution with ATP (159). These proteins were identified as BiP (grp78), HSP90 (grp94), calreticulin, a novel 46-kDa protein that binds azido-ATP, as well as three members of the thioredoxin superfamily: PDI, ERp72, and a previously reported 50-kDa protein (p50). Because the release of HSP90, PDI, ERp72, calreticulin, and p50 was stimulated by Ca^{2+}, these proteins appear to function as Ca^{2+}-dependent chaperones (159).

Evidence is accumulating that the ER HSP70 chaperone machinery is similar to that in the cytosol and bacteria, in that at least two DnaJ homologs have been found in the ER. For example, a yeast DnaJ homolog, Scj1p, is located in the lumen of the ER where it can interact with Kar2p (the HSP70 of the yeast ER) via the conserved J domain (198). Undoubtedly, chaperone-mediated folding in the lumen of the ER is complex, as revealed by the observation that the interaction of BiP with immunoglobulin light chains during folding suggests that light chains undergo both BiP-dependent and BiP-independent folding steps and that BiP must release the light chains before disulfide bond formation can occur in them (94).

B. Mitochondrial Import/Folding

Molecular chaperones play a critical role in targeting proteins to the mitochondria and the subsequent folding of the imported protein. In support of the endosymbiont theory on the origin of mitochondria, the chaperones of the mitochondria show a high degree of similarity to bacterial molecular chaperones, including a GrpE homolog (mGrpE) (193). The mitochondrial HSP70 (mHSP70) mediates protein transport across the inner membrane and protein folding in the matrix. These two reactions are carried out by two different mHSP70 complexes. The ADP-bound form of mHSP70 favors formation of a complex on the inner membrane; this "import complex" contains mHSP70, its membrane anchor Tim44, and mGrpE (106). The ATP-bound form of mHSP70 favors formation of a complex in the matrix; this "folding complex" contains mHSP70, the mitochondrial DnaJ homolog Mdj1, and mGrpE. A more detailed discussion of the role of chaperones in mitochondrial import and folding can be found in recent reviews (106, 156, 193).

V. MECHANISMS OF CHAPERONE FUNCTION

Considerable effort has been expended over the past few years to understand the mechanistic details of chaperone function. Great progress has been made, although considerable further study is necessary. The two best understood systems are those of HSP70 and GroEL. Even with these, the complexity of the systems, especially due to the interactions with cochaperones and other cofactors, has often led to apparently conflicting hypotheses. An additional source of potential discrepancies in behavior of the chaperones results from the effects of low concentrations of critical contaminants; for example, it has recently been shown that samples of HSP70 and HSP90 are often contaminated with low levels of DnaJ or HSP40, which may profoundly affect the experimental observations (204).

It is convenient to consider the mechanism of action of both HSP70 and GroEL in terms of their reaction cycles. Both of these chaperones require cochaperones for their full function, GroES in the case of GroEL, and HSP40 (or DnaJ) in the case of HSP70. Several theoretical models have been proposed to account for the effects of chaperonins on protein folding (25, 207, 232).

A. HSP70 Reaction Cycle

Several models have been proposed for the reaction cycle of HSP70 (4, 16, 74, 90, 146, 147, 166, 174, 195, 221). The DnaK cycle has been the most studied and is considered here. The reaction cycles for other HSP70 appear to be similar, with the exception that the cofactor GrpE will only be present in bacteria and mitochondria (273).

Although the general features of the HSP70 reaction cycle are established, there is considerable discussion about the details. Many observations indicate that the maximal functional effect of HSP70 requires the presence of DnaJ (or its homologs) (and GrpE in the case of prokaryotes and mitochondria) (73, 102, 136, 203, 217, 221, 249, 258, 274). The reports that DnaJ or HSP40 may bind at least some unfolded substrates are another source of confusion. It is now well established that GrpE and its homologs are nucleotide exchange factors and stimulate the ATPase cycle of DnaK or mHSP70 by increasing the rate of ADP release (39, 221) and that DnaJ and its homologs function to increase the rate of hydrolysis of HSP70-bound ATP (146, 221). Several studies have suggested that the action of DnaJ and GrpE with DnaK requires substoichiometric levels of the two cofactors (174). This is also consistent with the physiological molar ratios, in which DnaK is in large excess.

It has been shown that HSP70 discriminates between folded and unfolded proteins, normally binding only the latter (168). In fact, it is likely that HSP70 can distinguish between relatively unfolded intermediates and strongly nativelike intermediates and binds only the former. The fact that HSP70 binds to certain proteins in their native state, e.g., clathrin, is assumed to arise from the presence of accessible, unfolded loops. For a given unfolded substrate protein, there will be several potential HSP70 binding sites along the polypeptide chain, of different affinity for the chaperones (192). Both the conformational state of the substrate protein bound to HSP70 and the conformation of the substrate protein on ATP-induced release have been shown to be substantially unfolded (167).

The nature of the bound nucleotide affects the conformation of the chaperone and particularly its affinity for substrate protein. Thus complexes with ATP have low affinity for substrate and those with bound ADP have high affinity (166, 167, 174). The high affinity of HSP70 for nucleotides means that these chaperones will be found as binary complexes with ATP and ADP in the cell. Although the ATP complex binds substrate proteins/peptides much more rapidly than the ADP complex (146, 199, 224), the resulting ternary complex, HSP70-ATP-substrate, also releases the substrate protein very rapidly, and thus no productive complexes with unfolded substrate result (166). In contrast, the HSP70-ADP complex, although binding substrate protein at a slower rate, forms a relatively stable ternary complex, HSP70-ADP-substrate (167). The formation of small amounts of substrate complex when HSP70-ATP and substrate protein are mixed arises from ATP hydrolysis occurring during the reaction (stimulated by the presence of the substrate protein). Thus the formation of relatively long-lived complexes between unfolded proteins and HSP70 requires the presence of the HSP70-ADP-substrate complex. This explains, at least in part, the need for DnaJ and its homologs, since DnaJ significantly stimulates the rate of hydrolysis of ATP bound by HSP70, thus leading to formation of HSP70-ADP (18, 146).

The release of the substrate protein from the HSP70-ADP-substrate complex is triggered by the binding of ATP, which induces a conformational change in the peptide-binding domain (17, 84, 137, 165, 166, 227). On the basis of the crystallographic structure of the peptide-binding domain, the conformational change presumably involves the raising of the flap or latch, which is hypothesized to help maintain the substrate peptide bound (272). The cycle is completed by rebinding of another substrate protein molecule to the HSP70-ATP complex, or the hydrolysis of the ATP, leading to formation of DnaK-ADP and another conformational change. That substrate protein dissociation precedes ATP hydrolysis was demonstrated by comparison of the corresponding rates, the rate for ATP hydrolysis being significantly slower than that for substrate dissociation (166).

There appear to be several potential pathways for substrate proteins to enter the DnaK reaction cycle: via binding to DnaK-ATP, to DnaK-ATP-DnaJ, to DnaJ (which then binds to DnaK-ATP), to DnaK-ADP, and possibly to DnaK-ADP-DnaJ. The concentrations of the two ternary complexes are expected to be quite low so these are probably not major entry points. The same goes for DnaK-ADP under normal (nonstress) conditions when the levels of DnaK-ATP greatly exceed those of DnaK-ADP. The majority of the data suggest that the DnaK-ATP complex will normally be the main portal for entry to the cycle.

Most of the proposed reaction cycles of HSP70 fall into two broad classes: *1*) those which propose that it is only DnaK (HSP70) with bound ATP which interacts with the unfolded substrate, and that the interaction of this ternary complex with DnaJ (HSP40) leads to rapid ATP hydrolysis (146), and *2*) those which postulate that DnaJ first interacts with an unfolded (nascent) polypeptide, targeting it for binding to DnaK (74, 90, 221). Although there have been several reports that DnaJ (or HSP40) binds to some unfolded proteins (70, 95, 125, 126, 130, 203, 221), unambiguous evidence that DnaJ or its homologs will bind to unfolded proteins in general is currently lacking (240).

In the absence of the cochaperones, substrate protein will cycle on and off the ATP complex and accumulate only in the ADP complex. Although there are conflicting reports regarding the rate-limiting step in the intrinsic HSP70 ATPase activity, the evidence is strongly in favor of rate-limiting cleavage of the γ-phosphate of ATP, both in the absence and presence of DnaJ and substrate protein (115, 146, 147, 227). Both polypeptide substrates and DnaJ homologs stimulate the ATPase activity of HSP70 in *E. coli*, yeast, and human cytosol (147, 273). In the case of the yeast HSP70, Ssa1, the DnaJ homolog Ydj1 also accelerated release of ATP from Ssa1 (273), suggesting a possible explanation for the lack of a GrpE homolog in eukaryotic cytosol.

Thus the major pathway in the DnaK reaction cycle is likely to be the following (Fig. 4*A*). *1*) DnaK-ATP binds the unfolded substrate protein; the resulting complex may dissociate or bind DnaJ. *2*) The latter complex will undergo rapid DnaJ-stimulated hydrolysis of the ATP to yield a "stable" DnaK-ADP-substrate protein complex (due to the conformational change induced by the ATP→ADP transition), which may or may not also contain the DnaJ. *3*) The ADP dissociates, catalyzed by GrpE,

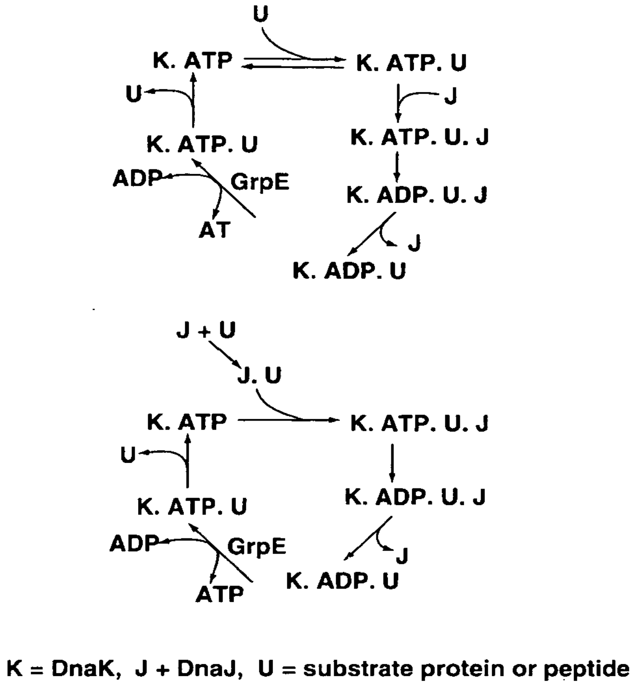


FIG. 4. Models of DnaK (HSP70) reaction cycle. *Top*: cycle starts with substrate protein binding to DnaK-ATP complex. *Bottom*: cycle starts with substrate protein binding to DnaJ. See text for details.

and is replaced by ATP. *4*) This induces a conformational change to the low-affinity form which results in dissociation of the substrate protein, leaving a DnaK-ATP complex. *5*) The latter can then either restart the cycle by binding a substrate protein, or it can undergo ATP hydrolysis to yield a DnaK-ADP complex. This would have to dissociate the ADP and rebind an ATP before entering the productive cycle again. The rates for several of the key steps in the DnaK cycle have been reported (4, 84, 117, 163, 174, 224). Because of the complexity of the system, the measured rates will be very sensitive to the concentrations of all the species involved, as well as the temperature and pH.

The alternative class of models in which DnaJ (or HSP40) acts as the initial chaperone will involve *1*) the unfolded substrate protein binding to DnaJ, which will then *2*) interact with DnaK-ATP, to form a transient HSP70-HSP40-U-ATP complex. This rapidly *3*) undergoes hydrolysis of its ATP, resulting in the formation of a stable HSP70-HSP40-U-ADP complex. It is likely that HSP40 dissociates rapidly from such a complex. Displacement of ADP by ATP (catalyzed by GrpE in bacteria and mitochondria) *4*) triggers the release of substrate protein, thus completing the reaction cycle (Fig. 4*B*).

Some of the newly synthesized proteins released by the HSP70 will fold spontaneously to the native state at a sufficiently fast rate that they neither aggregate nor bind to another chaperone molecule (either HSP70 or chaperonin) before they are fully folded. However, for some proteins, further interaction with a chaperonin, such as GroEL, is apparently required for complete folding (90, 96).

B. GroEL Reaction Cycle

The GroEL cycle is by far the most studied and best understood chaperonin reaction cycle, yet there are still outstanding questions. A comprehensive review has been published recently (53). For GroEL, the folding reaction is driven by cycles of binding and release of the cochaperone GroES, which alternate with binding and release of the nonnative protein substrate (62, 234). These cycles are driven by ATP binding and hydrolysis that control the conformation of the chaperonin and its affinity for nucleotides and the cochaperonin GroES. There are three major functional states: one in which the unfolded substrate is bound tightly, another in which the substrate protein is trapped in the cavity capped by GroES but in which folding can proceed because the substrate protein is not bound to the walls of the cavity, and a final state in which the substrate protein is "ejected" regardless of whether it is folded or not. Partially folded protein will rebind to the chaperonin, continuing the cycle until folding is complete (145). A distinction has been made between released nonnative conformations that are committed to folding and those that are not. It is assumed that the isolation of a partially folded intermediate in the GroES-capped, relatively polar cavity will lead to significant folding occurring, without competition from aggregation. Mutant chaperonins that are able to trap (bind but not release) substrate protein have proven very useful in such investigations (21, 52, 256).

Although both symmetric and asymmetric complexes of GroEL with GroES have been observed (211, 234, 237), only the latter are believed to be physiologically functional (194, 233) (although the existence of transient symmetric complexes cannot be ruled out). In the asymmetric complexes, the GroEL ring with GroES attached is known as the *cis*-ring, the opposing (distal) ring is the *trans*-ring. The recently determined structure of the GroEL-GroES-$(ADP)_7$ complex revealed that the large rigid-block movements of the intermediate and apical domains in the *cis*-ring allowed bound GroES to stabilize a folding chamber with ADP confined to the *cis*-ring (26, 188, 263). The conformational changes in the apical domains doubled the volume of the central cavity and resulted in burial of the hydrophobic peptide-binding residues at the interface with GroES and between the GroEL subunits. These structural changes result in the enlarged central cavity having a polar surface that favors protein folding (26, 263). The conformational changes induced in GroEL upon

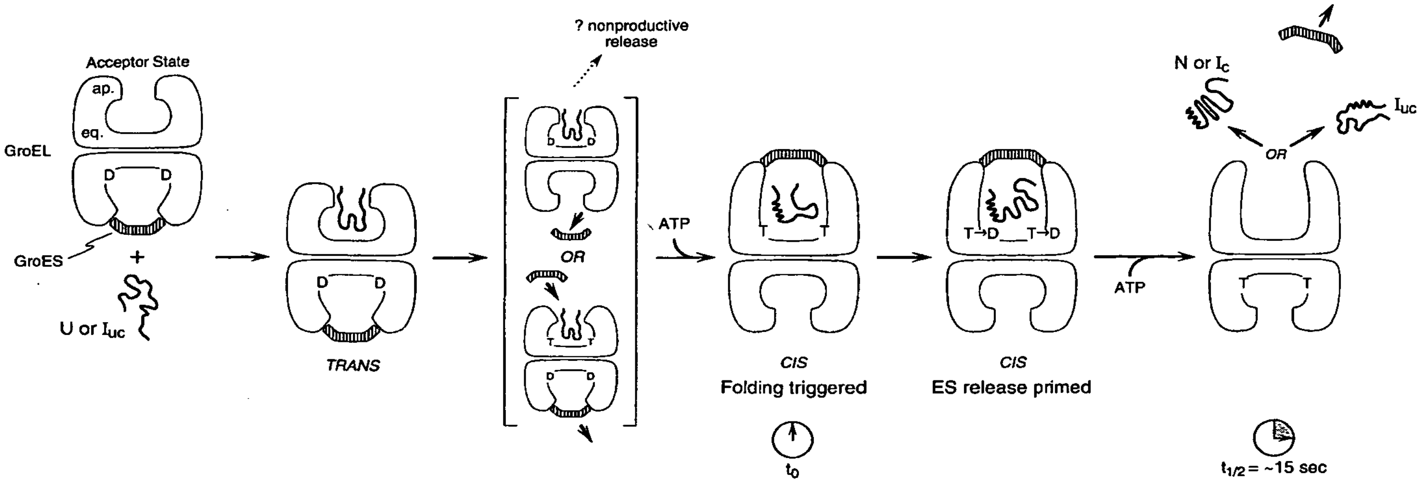


FIG. 5. GroEL reaction cycle. T and D represent ATP and ADP, respectively. I_c represents an intermediate committed to folding to native state, whereas I_{uc} represents intermediates that are not so committed. See text for details. [From Weissman et al. (251).]

binding of ATP have been observed by several techniques including electron microscopy imaging, X-ray crystallography, and fluorescence labeling (26, 140, 237, 263). In addition, it has been shown that binding of nucleotide to one GroEL ring is strongly favored by GroES binding to the other ring (211). The nucleotides bind to a site near the top of the equatorial domain facing the cavity (9). Under physiological concentrations of chaperonins (equimolar GroEL and GroES) and nucleotides, the predominant species is the asymmetric GroEL-GroES complex (20).

A recent model for the GroEL reaction cycle is shown in Figure 5. It is generally assumed that the asymmetric GroEL-GroES complex is the state that binds substrate protein (211). This species has ADP bound in the ring that is capped by GroES. The substrate protein thus binds to the hydrophobic cavity of the distal ring (20, 53, 145, 251). This *trans*-complex then converts to the *cis*-asymmetric complex in which GroES caps the cavity containing the substrate protein. This intermediate may arise either by a transient symmetric complex with two GroES lids, or by a complex in which the *trans*-GroES dissociates first. Binding of ATP in the *cis*-ring leads to the major conformational changes, especially in the apical domain leading to an increase in the size of the cavity and conversion of its surface to a more polar environment. This leads to dissociation of the substrate protein from its hydrophobic interactions with the lining of the cavity, favoring the folding reaction. Simultaneously, hydrolysis of the ATP in the *trans*-ring leads to the release of GroES and the opportunity for the substrate protein to exit the cavity. If the released substrate protein has not reached the native state, it may rebind for another cycle (53, 55).

Interactions between the two back-to-back rings in GroEL result in the allosteric regulation of ATP hydrolysis, binding, and release of folding substrates and the cochaperonin GroES. Allosterism in ATP hydrolysis can be described by a model in which each ring of GroEL is in equilibrium between a low-affinity (T) and high-affinity (R) state for ATP, and in which the GroEL double ring is in equilibrium between three states: TT, TR, and RR. Electron microscopy (26, 188) images of all three allosteric states, TT, TR, and RR, have been obtained for various complexes (26). Unfolded substrate proteins bind preferentially to the T state and stimulate the ATPase activity of GroEL by both a direct effect on GroEL and a shift in the equilibrium from the RR state toward the more active TR state (266). GroES promotes the T to R transition of the ring distal to GroES in the GroEL-GroES complex. Owing to the relatively low affinity of the R conformation for nonfolded proteins, this transition leads to release of protein substrates from *trans*-ternary complexes of GroEL, GroES, and protein substrate. The role of this release mechanism may be to assist the folding of relatively large proteins that cannot form *cis*-ternary complexes and/or to facilitate degradation of damaged proteins that cannot fold (111, 266). GroEL undergoes a conformational change that is partly maintained after ATP hydrolysis, as long as ADP and P_i are bound to the GroEL ring (140).

There have been several investigations of the rates for individual steps in the GroEL reaction cycle that demonstrate the importance of nucleotide binding and hydrolysis, and GroES binding, on the rate of substrate protein release (79, 91, 138, 139, 157, 182, 212, 234). Horwich and co-workers (21) have shown that under normal conditions the rate of hydrolysis of ATP in the ring *trans* to the bound GroES determines the rate of release of the GroES and hence the rate of dissociation of the substrate protein. This has been estimated to have a half-life of ~15 s (21). Confirmation that this "timer" sets the length of time for which the folding substrate protein remains in the GroEL cavity comes from observations on the folding of mitochondrial malate dehydrogenase; trapping experiments show that its dwell time on the complex is only 20 s (182). This is in good agreement with both the rate of ATP turnover and the dwell time of GroES on the complex but is much shorter than the time taken for the substrate to commit to the folded state.

Evidence is accumulating to indicate that GroEL is able to unfold misfolded conformations (1, 158, 183, 213, 232, 234, 268). The protection factors for the backbone amide protons of cyclophilin A bound to GroEL have been calculated from measurements of the rates of hydrogen/deuterium exchange using NMR (158); in contrast to the native structure, similar protection factors were found throughout the sequence consistent with complete unfolding of the substrate protein. Clarke and co-workers (183) studied the GroEL-facilitated folding of mitochondrial malate dehydrogenase and showed that the chaperonin accelerated the dissociation of a misfolded intermediate formed by reversible aggregation of an early partially folded intermediate, through a repeated binding and release cycle coupled to ATP hydrolysis. It is likely that the apparent "unfoldase" activity of GroEL actually arises from its preferential affinity for the unfolded conformation (247). Thus, through mass action, misfolded intermediates will be unfolded and given a new chance to fold productively in the GroEL cavity.

The key factors in the chaperonin cycle therefore are as follows: *1*) nonnative substrate protein binds to the *trans*-ring of GroEL, in which ADP and GroES are bound in the "opposite" GroEL ring. Binding is facilitated by the hydrophobic surfaces of the apical domain lining the cavity in the GroEL ring. *2*) Subsequent ATP binding to the *cis*-ring leads to release of the ADP and GroES, followed by *3*) binding of ATP and GroES to the *cis*-ring results in the massive conformational change leading to the enlarged cavity. This conformational change triggers the release of the substrate protein from the surface of the apical domain and also "starts the clock." *4*) Hydrolysis of the ATP in the *cis*-ring weakens the interaction between GroES and the *cis*-ring, and binding of ATP in the *trans*-ring leads to the complete release of the GroES and substrate protein with a half-life of ~15 s (194).

VI. CONCLUDING REMARKS

Molecular chaperones recognize and bind to nascent polypeptide chains and partially folded intermediates of proteins, preventing their aggregation and misfolding. The folding of most newly synthesized proteins in the cell will involve interaction with one or more chaperones. The chaperones most generally implicated in protein folding are the HSP40 (DnaJ), HSP60 (GroEL), and HSP70 (DnaK) families. Recent investigations using a wide variety of techniques ranging from genetics to biophysics have begun to unravel the complexities of these chaperone machines. At the heart of the general protein folding machinery of the cell are the reaction cycles of HSP60, HSP70, and their cochaperones. For both these chaperone systems, the binding of ATP triggers a critical conformational change ultimately leading to release of the bound substrate protein. Although both chaperone systems minimize aggregation of newly synthesized proteins, the HSP60 chaperones also facilitate the actual folding process by providing a secluded environment for individual folding molecules and may also promote the unfolding and refolding of misfolded intermediates. Different cellular locations, with their different roles in the production of new proteins, have specific chaperone systems tailored to the demands of the specific location (e.g., ER, mitochondria). Because of the critical nature of chaperones in maintaining orderly functioning of the cell, substantial redundancy is found in that multiple versions of chaperones are usually present. For selected proteins, additional specific chaperones are required for their folding and assembly. Although we now have what appears to be a good picture of the general outline of in vivo chaperone-mediated protein folding, it is clear that there are still a very large number of unanswered questions, especially regarding the molecular details.

REFERENCES

1. ACTON, S. L., D. H. WONG, P. PARHAM, F. M. BRODSKY, AND A. P. JACKSON. Alteration of clathrin light chain expression by transfection and gene disruption. *Mol. Biol. Cell* 4: 647–660, 1993.
2. ANFINSEN, C. B. Principles that govern the folding of protein chains. *Science* 181: 223–230, 1973.
3. ARTIGUES, A., A. IRIARTE, AND M. MARTINEZ-CARRION. Refolding intermediates of acid-unfolded mitochondrial aspartate aminotransferase bind to Hsp70. *J. Biol. Chem.* 272: 16852–16861, 1997.
4. BANECKI, B., AND M. ZYLICZ. Real time kinetics of the DnaK/DnaJ/GrpE molecular chaperone machine action. *J. Biol. Chem.* 271: 6137–6143, 1996.
5. BECKMANN, R. P., L. E. MIZZEN, AND W. J. WELCH. Interaction of Hsp 70 with newly synthesized proteins: implications for protein folding and assembly. *Science* 248: 850–854, 1990.
6. BERGERON, J. J., M. B. BRENNER, D. Y. THOMAS, AND D. B. WILLIAMS. Calnexin: a membrane-bound chaperone of the endoplasmic reticulum. *Trends Biochem. Sci.* 19: 124–128, 1994.
7. BLOND-ELGUINDI, S., S. E. CWIRLA, W. J. DOWER, R. J. LIPSHUTZ, S. R. SPRANG, J. F. SAMBROOK, AND M. J. GETHING. Affinity panning of a library of peptides displayed on bacteriophages reveals the binding specificity of BiP. *Cell* 75: 717–728, 1993.
8. BLUM, P., J. ORY, J. BAUERNFEIND, AND J. KRSKA. Physiological consequences of DnaK and DnaJ overproduction in *E. coli*. *J. Bacteriol.* 174: 7436–7444, 1992.
9. BOISVERT, D. C., J. WANG, Z. OTWINOWSKI, A. L. HORWICH, AND P. B. SIGLER. The 2.4 A crystal structure of the bacterial chaperonin GroEL complexed with ATP-γ-S. *Nature Struct. Biol.* 3: 170–177, 1996.
10. BOSE, S., T. WEIKL, H. BUGL, AND J. BUCHNER. Chaperone

function of Hsp90-associated proteins. *Science* 274: 1715–1717, 1996.

11. BOSTON, R. S., P. V. VIITANEN, AND E. VIERLING. Molecular chaperones and protein folding in plants. *Plant Mol. Biol.* 32: 191–222, 1996.
12. BRAIG, K., Z. OTWINOWSKI, R. HEGDE, D. C. BOISVERT, A. JOACHIMIAK, A. L. HORWICH, AND P. B. SIGLER. The crystal structure of the bacterial chaperonin GroEL at 2.8 A. *Nature* 371: 578–586, 1994.
13. BRAMHALL, E. A., R. L. CROSS, S. ROSPERT, N. K. STEEDE, AND S. J. LANDRY. Identification of amino acid residues at nucleotide-binding sites of chaperonin GroEL/GroES and cpn10 by photoaffinity labeling with 2-azido-adenosine 5′-triphosphate. *Eur. J. Biochem.* 244: 627–634, 1997.
14. BRODSKY, J. L. Translocation of proteins across the endoplasmic reticulum membrane. *Int. Rev. Cytol.* 178: 277–328, 1998.
15. BRUCE, B. D., AND J. CHURCHICH. Characterization of the molecular-chaperone function of the heat-shock-cognate-70-interacting protein. *Eur. J. Biochem.* 245: 738–744, 1997.
16. BUCHBERGER, A., H. SCHRODER, T. HESTERKAMP, H. J. SCHONFELD, AND B. BUKAU. Substrate shuttling between the DnaK and GroEL systems indicates a chaperone network promoting protein folding. *J. Mol. Biol.* 261: 328–333, 1996.
17. BUCHBERGER, A., H. THEYSSEN, H. SCHRODER, J. S. McCARTY, G. VIRGALLITA, P. MILKEREIT, J. REINSTEIN, AND B. BUKAU. Nucleotide-induced conformational changes in the ATPase and substrate binding domains of the DnaK chaperone provide evidence for interdomain communication. *J. Biol. Chem.* 270: 16903–16910, 1995.
18. BUCHBERGER, A., A. VALENCIA, R. McMACKEN, C. SANDER, AND B. BUKAU. The chaperone function of DnaK requires the coupling of ATPase activity with substrate binding through residue E171. *EMBO J.* 13: 1687–1695, 1994.
19. BUCHNER, J. Supervising the fold: functional principles of molecular chaperones. *FASEB J.* 10: 10–19, 1996.
20. BURSTON, S. G., AND A. R. CLARKE. Molecular chaperones: physical and mechanistic properties. *Essays Biochem.* 29: 125–136, 1995.
21. BURSTON, S. G., J. S. WEISSMAN, G. W. FARR, W. A. FENTON, AND A. L. HORWICH. Release of both native and non-native proteins from a *cis*-only GroEL ternary complex. *Nature* 383: 96–99, 1996.
22. CAI, H., C. C. WANG, AND C. L. TSOU. Chaperone-like activity of protein disulfide isomerase in the refolding of a protein with no disulfide bonds. *J. Biol. Chem.* 269: 24550–24552, 1994.
23. CAPLAN, A. J., D. M. CYR, AND M. G. DOUGLAS. Eukaryotic homologues of *Escherichia-coli* DnaJ: a diverse protein family that functions with Hsp70 stress proteins. *Mol. Biol. Cell* 4: 555–563, 1993.
24. CHAN, C. K., Y. HU, S. TAKAHASHI, D. L. ROUSSEAU, W. A. EATON, AND J. HOFRICHTER. Submillisecond protein folding kinetics studied by ultrarapid mixing. *Proc. Natl. Acad. Sci. USA* 94: 1779–1784, 1997.
25. CHAN, H. S., AND K. A. DILL. A simple model of chaperonin-mediated protein folding. *Proteins* 24: 345–351, 1996.
26. CHEN, S., A. M. ROSEMAN, A. S. HUNTER, S. P. WOOD, S. G. BURSTON, N. A. RANSON, A. R. CLARKE, AND H. R. SAIBIL. Location of a folding protein and shape changes in GroEL-GroES complexes imaged by cryo-electron microscopy. *Nature* 371: 261–264, 1994.
27. CHENG, M. Y., F. U. HARTL, J. MARTIN, R. A. POLLOCK, F. KALOUSEK, W. NEUPERT, E. M. HALLBERG, R. L. HALLBERG, AND A. L. HORWICH. Mitochondrial heat-shock protein Hsp60 is essential for assembly of proteins imported into yeast mitochondria. *Nature* 337: 620–625, 1989.
28. CLARKE, A. R., AND J. P. WALTHO. Protein folding and intermediates. *Curr. Opin. Biotechnol.* 8: 400–410, 1997.
29. CLARKE, J., L. S. ITZHAKI, AND A. R. FERSHT. Hydrogen exchange at equilibrium: a short cut for analysing protein- folding pathways? *Trends Biochem. Sci.* 22: 284–287, 1997.
30. COX, J. S., R. E. CHAPMAN, AND P. WALTER. The unfolded protein response coordinates the production of endoplasmic reticulum protein and endoplasmic reticulum membrane. *Mol. Biol. Cell* 8: 1805–1814, 1997.

30a. CREUTZ, C. E., A. LIOU, S. L. SNYDER, A. BROWNAWELL, AND K. WILLISON. Identification of the major chromaffin granule-binding protein, chromobindin A, as the cytosolic chaperonin CCT (chaperonin containing TCP-1). *J. Biol. Chem.* 269: 32035–32038, 1994.

31. CROOKE, E., AND W. WICKNER. Trigger factor: a soluble protein that folds pro-OmpA into a membrane. *Proc. Natl. Acad. Sci. USA* 84: 5216–5220, 1987.
32. CYR, D. M. Cooperation of the molecular chaperone Ydj1 with specific Hsp70 homologs to suppress protein aggregation. *FEBS Lett.* 359: 129–132, 1995.
33. CYR, D. M., AND M. G. DOUGLAS. Differential regulation of Hsp70 subfamilies by the eukaryotic DnaJ homologue Ydj1. *J. Biol. Chem.* 269: 9798–9804, 1994.
34. CYR, D. M., T. LANGER, AND M. G. DOUGLAS. DnaJ-like proteins: molecular chaperones and specific regulators of Hsp70. *Trends Biochem. Sci.* 19: 176–181, 1994.
35. CYR, D. M., X. LU, AND M. G. DOUGLAS. Regulation of Hsp70 function by a eukaryotic DnaJ homolog. *J. Biol. Chem.* 267: 20927–20931, 1992.
36. DARBY, N., AND T. E. CREIGHTON. Probing protein folding and stability using disulfide bonds. *Mol. Biotechnol.* 7: 57–77, 1997.
37. DEBBURMAN, S. K., G. J. RAYMOND, B. CAUGHEY, AND S. LINDQUIST. Chaperone-supervised conversion of prion protein to its protease-resistant form. *Proc. Natl. Acad. Sci. USA* 94: 13938–13943, 1997
38. DE CROUY-CHANEL, A., M. KOHIYAMA, AND G. RICHARME. A novel function of *Escherichia coli* chaperone DnaJ. Protein-disulfide isomerase. *J. Biol. Chem.* 270: 22669–22672, 1995.
39. DEKKER, P. J., AND N. PFANNER. Role of mitochondrial GrpE and phosphate in the ATPase cycle of matrix Hsp70. *J. Mol. Biol.* 270: 321–327, 1997.
40. DELOCHE, O., W. L. KELLEY, AND C. GEORGOPOULOS. Structure-function analyses of the Ssc1p, Mdj1p, and Mge1p *Saccharomyces cerevisiae* mitochondrial proteins in *Escherichia coli*. *J. Bacteriol.* 179: 6066–6075, 1997.
41. DELOCHE, O., K. LIBEREK, M. ZYLICZ, AND C. GEORGOPOULOS. Purification and biochemical properties of *Saccharomyces cerevisiae* Mdj1p, the mitochondrial DnaJ homologue. *J. Biol. Chem.* 272: 28539–28544, 1997.
42. DEYOUNG, L. R., A. L. FINK, AND K. A. DILL. Aggregation of globular proteins. *Acc. Chem. Res.* 26: 614–620, 1993.
43. DILL, K. A., S. BROMBERG, K. Z. YUE, K. M. FIEBIG, D. P. YEE, P. D. THOMAS, AND H. S. CHAN. Principles of protein folding: a perspective from simple exact models. *Protein Sci.* 4: 561–602, 1995.
44. DILL, K. A., AND H. S. CHAN. From Levinthal to pathways to funnels. *Nature Struct. Biol.* 4: 10–19, 1997.
45. DITTMAR, K. D., AND W. B. PRATT. Folding of the glucocorticoid receptor by the reconstituted Hsp90-based chaperone machinery. The initial Hsp90.p60.Hsp70-dependent step is sufficient for creating the steroid binding conformation. *J. Biol. Chem.* 272: 13047–13054, 1997.
46. EGGERS, D. K., W. J. WELCH, AND W. J. HANSEN. Complexes between nascent polypeptides and their molecular chaperones in the cytosol of mammalian cells. *Mol. Biol. Cell* 8: 1559–1573, 1997.
47. EHRNSPERGER, M., M. GAESTEL, AND J. BUCHNER. Structure and function of small heat-shock proteins. In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto. New York: Dekker, 1998, p. 533–575.
48. ELLIS, R. J. The molecular chaperone concept. *Semin. Cell Biol.* 1: 1–9, 1990.
49. ELLIS, R. J. Revisiting the Anfinsen cage. *Folding Design* 1: R9—R15, 1996.
50. ELLIS, R. J., AND F. U. HARTL. Protein folding in the cell: competing models of chaperonin function. *FASEB J.* 10: 20–26, 1996.
51. ENGLANDER, S. W., L. MAYNE, Y. BAI, AND T. R. SOSNICK. Hydrogen exchange: the modern legacy of Linderstrom-Lang. *Protein Sci.* 6: 1101–1109, 1997.
52. FARR, G. W., E. C. SCHARL, R. J. SCHUMACHER, S. SONDEK, AND A. L. HORWICH. Chaperonin-mediated folding in the eukaryotic cytosol proceeds through rounds of release of native and nonnative forms. *Cell* 89: 927–937, 1997.

53. FENTON, W. A., AND A. L. HORWICH. GroEL-mediated protein folding. *Protein Sci.* 6: 743–760, 1997.
54. FENTON, W. A., Y. KASHI, K. FURTAK, AND A. L. HORWICH. Residues in chaperonin GroEL required for polypeptide binding and release. *Nature* 371: 614–619, 1994.
55. FENTON, W. A., J. S. WEISSMAN, AND A. L. HORWICH. Putting a lid on protein folding: structure and function of the co-chaperonin, GroES. *Chem. Biol.* 3: 157–161, 1996.
56. FERSHT, A. R. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* 7: 3–9, 1997.
57. FINK, A. L. Compact intermediate states in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* 24: 495–522, 1995.
58. FINK, A. L. The Hsp 70 reaction cycle and its role in protein folding. In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto New York: Dekker, 1998, p. 123–150.
59. FINK, A. L. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Folding Design* 3: R9—R15, 1998.
60. FINK, A. L., AND Y. GOTO (Editors). *Molecular Chaperones in the Life Cycle of Proteins.* New York: Dekker, 1998.
61. FINK, A. L., K. A. OBERG, AND S. SESHADRI. Discrete intermediates vs. molten globule models of protein folding: characterization of partially-folded intermediates of apomyoglobin. *Folding Design* 3: 19–25, 1997.
62. FISHER, M. T., AND X. YUAN. The rates of commitment to renaturation of rhodanese and glutamine synthetase in the presence of the GroE chaperonins. *J. Biol. Chem.* 269: 29598–29601, 1994.
63. FLAHERTY, K. M., C. DELUCA-FLAHERTY, AND D. B. MCKAY. Three-dimensional structure of the ATPase fragment of a 70K heat-shock protein. *Nature* 346: 623–628, 1990.
64. FLYNN, G. C., J. POHL, M. T. FLOCCO, AND J. E. ROTHMAN. Peptide-binding specificity of the molecular chaperone BiP. *Nature* 353: 726–730, 1991.
65. FREEDMAN, R. B., T. R. HIRST, AND M. F. TUITE. Protein disulphide isomerase: building bridges in protein folding. *Trends Biochem. Sci.* 19: 331–336, 1994.
66. FREEMAN, B. C., AND R. I. MORIMOTO. The human cytosolic molecular chaperones Hsp90, Hsp70 (Hsc70) and Hdj-1 have distinct roles in recognition of a non-native protein and protein refolding. *EMBO J.* 15: 2969–2979, 1996.
67. FREEMAN, B. C., D. O. TOFT, AND R. I. MORIMOTO. Molecular chaperone machines: chaperone activities of the cyclophilin Cyp-40 and the steroid aporeceptor-associated protein p23. *Science* 274: 1718–1720, 1996.
68. FRYDMAN, J., AND F. U. HARTL. Principles of chaperone-assisted protein folding: differences between in vitro and in vivo mechanisms. *Science* 272: 1497–1502, 1996.
69. FRYDMAN, J., E. NIMMESGERN, H. ERDJUMENT-BROMAGE, J. S. WALL, P. TEMPST, AND F.-U. HARTL. Function in protein folding of TriC, a cytosolic ring complex containing tcp-1 and structurally related subunits. *EMBO J.* 11: 4767–4778, 1992.
70. FRYDMAN, J., E. NIMMESGERN, K. OHTSUKA, AND F. U. HARTL. Folding of nascent polypeptide chains in a high molecular mass assembly with molecular chaperones. *Nature* 370: 111–117, 1994.
71. FUNG, K. L., L. HILGENBERG, N. M. WANG, AND W. J. CHIRICO. Conformations of the nucleotide and polypeptide binding domains of a cytosolic Hsp70 molecular chaperone are coupled. *J. Biol. Chem.* 271: 21559–21565, 1996.
72. GAITANARIS, G. A., A. VYSOKANOV, S. C. HUNG, M. E. GOTTESMAN, AND A. GRAGEROV. Successive action of *Escherichia coli* chaperones in vivo. *Mol. Microbiol.* 14: 861–869, 1994.
73. GAMER, J., H. BUJARD, AND B. BUKAU. Physical interaction between heat shock proteins DnaK, DnaJ, and GrpE and the bacterial heat shock transcription factor-sigma(32). *Cell* 69: 833–842, 1992.
74. GAMER, J., G. MULTHAUP, T. TOMOYASU, J. S. MCCARTY, S. RUDIGER, H. J. SCHONFELD, C. SCHIRRA, H. BUJARD, AND B. BUKAU. A cycle of binding and release of the DnaK, DnaJ and GrpE chaperones regulates activity of the *Escherichia coli* heat shock transcription factor sigma32. *EMBO J.* 15: 607–617, 1996.
75. GEORGOPOULOS, C., AND W. J. WELCH. Role of the major heat shock proteins as molecular chaperones. *Annu. Rev. Cell Biol.* 9: 601–634, 1993.
76. GETHING, M. J., AND J. SAMBROOK. Protein folding in the cell. *Nature* 355: 33–45, 1992.
77. GILBERT, H. F. Protein disulfide isomerase and assisted protein folding. *J. Biol. Chem.* 272: 29399–29402, 1997.
78. GLOVER, J. R., E. C. SCHIRMER, M. A. SINGER, AND S. L. LINDQUIST. Hsp104. In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto. New York: Dekker, 1998, p. 193–224.
79. GOLDBERG, M. S., J. ZHANG, S. SONDEK, C. R. MATTHEWS, R. O. FOX, AND A. L. HORWICH. Native-like structure of a protein-folding intermediate bound to the chaperonin GroEL. *Proc. Natl. Acad. Sci. USA* 94: 1080–1085, 1997.
80. GRAGEROV, A., AND M. E. GOTTESMAN. Different peptide binding specificities of Hsp70 family members. *J. Mol. Biol.* 241: 133–135, 1994.
81. GRAGEROV, A., E. NUDLER, N. KOMISSAROVA, G. A. GAITANARIS, M. E. GOTTESMAN, AND V. NIKIFOROV. Cooperation of GroEL/GroES and DnaK/DnaJ heat shock proteins in preventing protein misfolding in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 89: 10341–10344, 1992.
82. GRENERT, J. P., W. P. SULLIVAN, P. FADDEN, T. A. J. HAYSTEAD, J. CLARK, E. MIMNAUGH, H. KRUTZSCH, H. J. OCHEL, T. W. SCHULTE, E. SAUSVILLE, L. M. NECKERS, AND D. O. TOFT. The amino-terminal domain of heat shock protein 90 (hsp90) that binds geldanamycin is an ATP/ADP switch domain that regulates hsp90 conformation. *J. Biol. Chem.* 272: 23843–23850, 1997.
83. HA, J.-H., E. R. JOHNSON, D. B. MCKAY, M. C. SOUSA, S. TAKEDA, AND S. M. WILBANKS. Structure and properties of the 70-kilodalton heat-shock proteins. In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto. New York: Dekker, 1998, p. 95–122.
84. HA, J. H., AND D. B. MCKAY. Kinetics of nucleotide-induced changes in the tryptophan fluorescence of the molecular chaperone Hsc70 and its subfragments suggest the ATP-induced conformational change follows initial ATP binding. *Biochemistry* 34: 11635–11644, 1995.
85. HAAS, I. G., AND M. WABL. Immunoglobulin heavy chain binding protein. *Nature* 306: 387–389, 1983.
86. HANSEN, W. J., V. R. LINGAPPA, AND W. J. WELCH. Complex environment of nascent polypeptide chains. *J. Biol. Chem.* 269: 26610–26613, 1994.
87. HARDESTY, B., W. KUDLICKI, O. W. ODOM, T. ZHANG, D. McCARTHY, AND G. KRAMER. Cotranslational folding of nascent proteins on *Escherichia coli* ribosomes. *Biochem. Cell Biol.* 73: 1199–1207, 1995.
88. HARDY, S. J., AND L. L. RANDALL. Recognition of ligands by SecB, a molecular chaperone involved in bacterial protein export. *Philos. Trans. R Soc. Lond. B Biol. Sci.* 339: 343–354, 1993.
89. HARRISON, C. J., M. HAYER-HARTL, M. DI LIBERTO, F. HARTL, AND J. KURIYAN. Crystal structure of the nucleotide exchange factor GrpE bound to the ATPase domain of the molecular chaperone DnaK. *Science* 276: 431–435, 1997.
90. HARTL, F. U. Molecular chaperones in cellular protein folding. *Nature* 381: 571–579, 1996.
91. HAYER-HARTL, M. K., J. MARTIN, AND F. U. HARTL. Asymmetrical interaction of GroEL and GroES in the ATPase cycle of assisted protein folding. *Science* 269: 836–841, 1995.
92. HEBERT, D. N., B. FOELLMER, AND A. HELENIUS. Glucose trimming and reglucosylation determine glycoprotein association with calnexin in the endoplasmic reticulum. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 348: 107–112, 1995.
93. HEBERT, D. N., J. F. SIMONS, J. R. PETERSON, AND A. HELENIUS. Calnexin, calreticulin, and Bip/Kar2p in protein folding. *Cold Spring Harb. Symp. Quant. Biol.* 60: 405–415, 1995.
94. HENDERSHOT, L., J. WEI, J. GAUT, J. MELNICK, S. AVIEL, AND Y. ARGON. Inhibition of immunoglobulin folding and secretion by dominant negative BiP ATPase mutants. *Proc. Natl. Acad. Sci. USA* 93: 5269–5274, 1996.
95. HENDRICK, J. P., AND F. U. HARTL. Molecular chaperone functions of heat-shock proteins. *Annu. Rev. Biochem.* 62: 349–384, 1993.
96. HENDRICK, J. P., AND F. U. HARTL. The role of molecular chaperones in protein folding. *FASEB J.* 9: 1559–1569, 1995.
97. HENDRICK, J. P., T. LANGER, T. A. DAVIS, F. U. HARTL, AND M. WIEDMANN. Control of folding and membrane translocation by

binding of the chaperone DnaJ to nascent polypeptides. *Proc. Natl. Acad. Sci. USA* 90: 10216–10220, 1993.

98. HESTERKAMP, T., AND B. BUKAU. The *Escherichia coli* trigger factor. *FEBS Lett.* 389: 32–34, 1996.

99. HESTERKAMP, T., S. HAUSER, H. LUTCKE, AND B. BUKAU. *Escherichia coli* trigger factor is a prolyl isomerase that associates with nascent polypeptide chains. *Proc. Natl. Acad. Sci. USA* 93: 4437–4441, 1996.

100. HIGHTOWER, L. E., AND S.-M. LEUNG. Substrate-binding specificty of the Hsp70 family. In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto. New York: Dekker, 1998, p. 151–168.

101. HILL R. B., J. M. FLANAGAN, AND J. H. PRESTEGARD. ^{1}H and ^{15}N magnetic resonance assignments, secondary structure, and tertiary fold of *Escherichia coli* DnaJ(1–78). *Biochemistry* 34: 5587–5596, 1995.

102. HOFFMANN, H. J., S. K. LYMAN, C. LU, M. A. PETIT, AND H. ECHOLS. Activity of the Hsp70 chaperone complex–DnaK, DnaJ, and GrpE–in initiating phage lambda DNA replication by sequestering and releasing lambda P protein. *Proc. Natl. Acad. Sci. USA* 89: 12108–12111, 1992.

103. HOHFELD, J., AND S. JENTSCH. GrpE-like regulation of the Hsc70 chaperone by the anti-apoptotic protein BAG-1. *EMBO J.* 16: 6209–6216, 1997.

104. HOHFELD, J., Y. MINAMI, AND F. U. HARTL. Hip, a novel cochaperone involved in the eukaryotic Hsc70/Hsp40 reaction cycle. *Cell* 83: 589–598, 1995.

105. HOROWITZ, P. M. Some structural aspects of chaperonin-assisted folding by GroEL and GroES. In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto. New York: Dekker, 1998, p. 275–300.

106. HORST, M., W. OPPLIGER, S. ROSPERT, H. J. SCHONFELD, G. SCHATZ, AND A. AZEM. Sequential action of two Hsp70 complexes during protein import into mitochondria. *EMBO J.* 16: 1842–1849, 1997.

107. HORWICH, A. L., K. B. LOW, W. A. FENTON, I. N. HIRSHFIELD, AND K. FURTAK. Folding in vivo of bacterial cytoplasmic proteins: role of GroEL. *Cell* 74: 909–917, 1993.

108. HORWICH, A. L., AND K. R. WILLISON. Protein folding in the cell: functions of two families of molecular chaperone, Hsp 60 and TF55-TCP1. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 339: 313–325, 1993.

109. HU, G., T. GURA, B. SABSAY, J. SAUK, S. N. DIXIT, AND A. VEIS. Endoplasmic reticulum protein Hsp47 binds specifically to the N-terminal globular domain of the amino-propeptide of the procollagen I alpha 1 (I)-chain. *J. Cell Biochem.* 59: 350–367, 1995.

110. HUNT, J. F., A. J. WEAVER, S. J. LANDRY, L. GIERASCH, AND J. DEISENHOFER. The crystal structure of the GroES co-chaperonin at 2.8 A resolution. *Nature* 379: 37–45, 1996.

111. INBAR, E., AND A. HOROVITZ. GroES promotes the T to R transition of the GroEL ring distal to GroES in the GroEL-GroES complex. *Biochemistry* 36: 12276–12281, 1997.

112. JAENICKE, R. Folding and association versus misfolding and aggregation of proteins. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 348: 97–105, 1995.

113. JAKOB, U., AND J. BUCHNER. Assisting spontaneity: the role of Hsp90 and small Hsps as molecular chaperones. *Trends Biochem. Sci.* 19: 205–211, 1994.

114. JOHNSON, B. D., R. J. SCHUMACHER, E. D. ROSS, AND D. O. TOFT. Hop modulates Hsp70/Hsp90 interactions in protein folding. *J. Biol. Chem.* 273: 3679–3686, 1998.

115. KAMATH-LOEB, A. S., C. Z. LU, W. C. SUH, M. A. LONETTO, AND C. A. GROSS. Analysis of three DnaK mutant proteins suggests that progression through the ATPase cycle requires conformational changes. *J. Biol. Chem.* 270: 30051–30059, 1995.

116. KANDROR, O., M. SHERMAN, R. MOERSCHELL, AND A. L. GOLDBERG. Trigger factor associates with GroEL in vivo and promotes its binding to certain polypeptides. *J. Biol. Chem.* 272: 1730–1734, 1997.

117. KARZAI, A. W., AND R. McMACKEN. A bipartite signaling mechanism involved in DnaJ-mediated activation of the *Escherichia coli* DnaK protein. *J. Biol. Chem.* 271: 11236–11246, 1996.

118. KIMURA, Y., I. YAHARA, AND S. LINDQUIST. Role of the protein chaperone YDJ1 in establishing Hsp90-mediated signal transduction pathways. *Science* 268: 1362–1365, 1995.

119. KRAUSE, K. H., AND M. MICHALAK. Calreticulin. *Cell* 88: 439–443, 1997.

121. KUBOTA, H., G. HYNES, A. CARNE, A. ASHWORTH, AND K. WILLISON. Identification of six Tcp-1-related genes encoding divergent subunits of the TCP-1-containing chaperonin. *Curr. Biol.* 4: 89–99, 1994.

122. KUBOTA, H., G. HYNES, AND K. WILLISON. The chaperonin containing t-complex polypeptide 1 (TCP-1). Multisubunit machinery assisting in protein folding and assembly in the eukaryotic cytosol. *Eur. J. Biochem.* 230: 3–16, 1995.

123. KUDLICKI, W., J. CHIRGWIN, G. KRAMER, AND B. HARDESTY. Folding of an enzyme into an active conformation while bound as peptidyl-tRNA to the ribosome. *Biochemistry* 34: 14284–14287, 1995.

124. KUDLICKI, W., O. W. ODOM, G. KRAMER, AND B. HARDESTY. Chaperone-dependent folding and activation of ribosome-bound nascent rhodanese. Analysis by fluorescence. *J. Mol. Biol.* 244: 319–331, 1994.

125. KUDLICKI, W., O. W. ODOM, G. KRAMER, AND B. HARDESTY. Binding of an N-terminal rhodanese peptide to DnaJ and to ribosomes. *J. Biol. Chem.* 271: 31160–31165, 1996.

126. KUDLICKI, W., O. W. ODOM, G. KRAMER, B. HARDESTY, G. A. MERRILL, AND P. M. HOROWITZ. The importance of the N-terminal segment for DnaJ-mediated folding of rhodanese while bound to ribosomes as peptidyl-tRNA. *J. Biol. Chem.* 270: 10650–10657, 1995.

127. KUEHN, M. J., D. J. OGG, J. KIHLBERG, L. N. SLONIM, K. FLEMMER, T. BERGFORS, AND S. J. HULTGREN. Structural basis of pilus subunit recognition by the PapD chaperone. *Science* 262: 1234–1241, 1993.

128. LANDRY, S. J., A. TAHER, C. GEORGOPOULOS, AND S. M. VAN DER VIES. Interplay of structure and disorder in cochaperonin mobile loops. *Proc. Natl. Acad. Sci. USA* 93: 11622–11627, 1996.

129. LANDRY, S. J., J. ZEILSTRA-RYALLS, O. FAYET, C. GEORGOPOULOS, AND L. M. GIERASCH. Characterization of a functionally important mobile domain of GroES. *Nature* 364: 255–258, 1993.

130. LANGER, T., C. LU, H. ECHOLS, J. FLANAGAN, M. K. HAYER, AND F.-U. HARTL. Successive action of DnaK, DnaJ and GroEL along the pathway of chaperone-mediated protein folding. *Nature* 356: 683–689, 1992.

131. LAUFEN, T., U. ZUBER, A. BUCHBERGER, AND B. BUKAU. DnaJ proteins In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto. New York: Dekker, 1998, p. 241–274.

132. LEE, D. H., M. Y. SHERMAN, AND A. L. GOLDBERG. Involvement of the molecular chaperone Ydj1 in the ubiquitin-dependent degradation of short-lived and abnormal proteins in *Saccharomyces cerevisiae. Mol. Cell. Biol.* 16: 4773–4781, 1996.

133. LEE, G. J., A. M. ROSEMAN, H. R. SAIBIL, AND E. VIERLING. A small heat shock protein stably binds heat-denatured model substrates and can maintain a substrate in a folding-competent state. *EMBO J.* 16: 659–671, 1997.

134. LEROUX, M. R., R. MELKI, B. GORDON, G. BATELIER, AND E. P. CANDIDO. Structure-function studies on small heat shock protein oligomeric assembly and interaction with unfolded polypeptides. *J. Biol. Chem.* 272: 24646–24656, 1997.

135. LEUNG, S. M., AND L. E. HIGHTOWER. A 16-kDa protein functions as a new regulatory protein for Hsc70 molecular chaperone and is identified as a member of the Nm23/nucleoside diphosphate kinase family. *J. Biol. Chem.* 272: 2607–2614, 1997.

136. LIBEREK, K., C. GEORGOPOULOS, AND M. ZYLICZ. Role of the *Escherichia coli* DnaK and DnaJ heat shock proteins in the initiation of bacteriophage lambda DNA replication. *Proc. Natl. Acad. Sci. USA* 85: 6632–6636, 1988.

137. LIBEREK, K., J. MARSZALEK, D. ANG, C. GEORGOPOULOS, AND M. ZYLICZ. *Escherichia coli* DnaJ and GrpE heat shock proteins jointly stimulate ATPase activity of DnaK. *Proc. Natl. Acad. Sci. USA* 88: 2874–2878, 1991.

138. LILIE, H., AND J. BUCHNER. Interaction of GroEL with a highly structured folding intermediate: iterative binding cycles do not involve unfolding. *Proc. Natl. Acad. Sci. USA* 92: 8100–8104, 1995.

139. LIN, Z., AND E. EISENSTEIN. Nucleotide binding-promoted conformational changes release a nonnative polypeptide from the *Escherichia coli* chaperonin GroEL. *Proc. Natl. Acad. Sci. USA* 93: 1977–1981, 1996.

140. LLORCA, O., S. MARCO, J. L. CARRASCOSA, AND J. M. VALPUESTA. Conformational changes in the GroEL oligomer during the functional cycle. *J. Struct. Biol.* 118: 31–42, 1997.

141. LLOSA, M., K. ALORIA, R. CAMPO, R. PADILLA, J. AVILA, L. SANCHEZ-PULIDO, AND J. C. ZABALA. The beta-tubulin monomer release factor (p14) has homology with a region of the DnaJ protein. *FEBS Lett.* 397: 283–289, 1996.

142. LORIMER, G. H. A quantitative assessment of the role of the chaperonin proteins in protein folding in vivo. *FASEB J.* 10: 5–9, 1996.

143. LUZ, J. M., AND W. J. LENNARZ. Protein disulfide isomerase: a multifunctional protein of the endoplasmic reticulum. *EXS* 77: 97–117, 1996.

144. MARTIN, J., AND F. U. HARTL. Chaperone-assisted protein folding. *Curr. Opin. Struct. Biol.* 7: 41–52, 1997.

145. MARTIN, J., M. MAYHEW, T. LANGER, AND F. U. HARTL. The reaction cycle of GroEL and GroES in chaperonin-assisted protein folding. *Nature* 366: 228–233, 1993.

146. McCARTY, J. S., A. BUCHBERGER, J. REINSTEIN, AND B. BUKAU. The role of ATP in the functional cycle of the DnaK chaperone system. *J. Mol. Biol.* 249: 126–137, 1995.

147. MINAMI, Y., J. HOHFELD, K. OHTSUKA, AND F. U. HARTL. Regulation of the heat-shock protein 70 reaction cycle by the mammalian DnaJ homolog, Hsp40. *J. Biol. Chem.* 271: 19617–19624, 1996.

148. MITRAKI, A., B. FANE, C. HAASE-PETTINGELL, J. STURTEVANT, AND J. KING. Global suppression of protein folding defects and inclusion body formation. *Science* 253: 54–58, 1991.

149. MORI, K., T. KAWAHARA, H. YOSHIDA, H. YANAGI, AND T. YURA. Signalling from endoplasmic reticulum to nucleus: transcription factor with a basic-leucine zipper motif is required for the unfolded protein-response pathway. *Genes Cells* 1: 803–817, 1996.

150. MORIMOTO, R. I., A. TISSIERES, AND C. GEORGOPOULOS (Editors). *Progress and Perspectives on the Biology of Heat Shock Proteins and Molecular Chaperones.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 1994, p. 1–30.

151. MUNRO, S., AND H. R. PELHAM. An Hsp70-like protein in the ER: identity with the 78 kd glucose-regulated protein and immunoglobulin heavy chain binding protein. *Cell* 46: 291–300, 1986.

152. NAGATA, K., M. SATOH, A. D. MILLER, AND N. HOSOKAWA. Involvement of Hsp47 in the folding and processing of procollagen in the endoplasmic reticulum. In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto. New York: Dekker, 1998, p. 225–240.

153. NATHAN, D. F., M. H. VOS, AND S. LINDQUIST. In vivo functions of the *Saccharomyces cerevisiae* Hsp90 chaperone. *Proc. Natl. Acad. Sci. USA* 94: 12949–12956, 1997.

154. NELSON, R. J., T. ZIEGELHOFFER, C. NICOLET, M. WERNER-WASHBURNE, AND E. A. CRAIG. The translation machinery and 70 kd heat shock protein cooperate in protein synthesis. *Cell* 71: 97–105, 1992.

155. NETZER, W. J., AND F. U. HARTL. Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* 388: 343–349, 1997.

156. NEUPERT, W. Protein import into mitochondria. *Annu. Rev. Biochem.* 66: 863–917, 1997.

157. NIEBA, L., S. E. NIEBA-AXMANN, A. PERSSON, M. HAMALAINEN, F. EDEBRATT, A. HANSSON, J. LIDHOLM, K. MAGNUSSON, A. F. KARLSSON, AND A. PLUCKTHUN. BIACORE analysis of histidine-tagged proteins using a chelating NTA sensor chip. *Anal. Biochem.* 252: 217–228, 1997.

158. NIEBA-AXMANN, S. E., M. OTTIGER, K. WUTHRICH, AND A. PLUCKTHUN. Multiple cycles of global unfolding of GroEL-bound cyclophilin A evidenced by NMR. *J. Mol. Biol.* 271: 803–818, 1997.

159. NIGAM, S. K., A. L. GOLDBERG, S. HO, M. F. ROHDE, K. T. BUSH, AND M. Y. U. SHERMAN. A set of endoplasmic reticulum proteins possessing properties of molecular chaperones includes Ca^{2+}-binding proteins and members of the thioredoxin superfamily. *J. Biol. Chem.* 269: 1744–1749, 1994.

160. NIMMESGERN, E., AND F. U. HARTL. ATP-dependent protein refolding activity in reticulocyte lysate. Evidence for the participation of different chaperone components. *FEBS Lett.* 331: 25–30, 1993.

161. OBERMOELLER, L. M., I. WARSHAWSKY, M. R. WARDELL, G. BU, AND R. RAPPUOLI. Differential functions of triplicated repeats suggest two independent roles for the receptor-associated protein as a molecular chaperone. Efficient production of heat-labile enterotoxin mutant proteins by overexpression of dsbA in a degP-deficient *Escherichia coli* strain. *Arch. Microbiol.* 167: 280–283, 1997.

162. OU, W. J., P. H. CAMERON, D. Y. THOMAS, AND J. J. BERGERON. Association of folding intermediates of glycoproteins with calnexin during protein maturation. *Virology* 193: 545–562, 1993.

163. PACKSCHIES, L., H. THEYSSEN, A. BUCHBERGER, B. BUKAU, R. S. GOODY, AND J. REINSTEIN. GrpE accelerates nucleotide exchange of the molecular chaperone DnaK with an associative displacement mechanism. *Biochemistry* 36: 3417–3422, 1997.

164. PAHL, A., K. BRUNE, AND H. BANG. Fit for life? Evolution of chaperones and folding catalysts parallels the development of complex organisms. *Cell Stress Chaperones* 2: 78–86, 1997.

165. PALLEROS, D. R., K. L. REID, J. S. McCARTY, G. C. WALKER, AND A. L. FINK. Dnak, Hsp73, and their molten globules. Two different ways heat shock proteins respond to heat. *J. Biol. Chem.* 267: 5279–5285, 1992.

166. PALLEROS, D. R., K. L. REID, L. SHI, W. J. WELCH, AND A. L. FINK. ATP-induced protein Hsp70 complex dissociation requires K^+ but not ATP hydrolysis. *Nature* 365: 664–666, 1993.

167. PALLEROS, D. R., L. SHI, K. L. REID, AND A. L. FINK. Hsp70-protein complexes. Complex stability and conformation of bound substrate protein. *J. Biol. Chem.* 269: 13107–13114, 1994.

168. PALLEROS, D. R., W. J. WELCH, AND A. L. FINK. Interaction of hsp70 with unfolded proteins: effects of temperature and nucleotides on the kinetics of binding. *Proc. Natl. Acad. Sci. USA* 88: 5719–5723, 1991.

169. PANAGIOTIDIS, C. A., W. F. BURKHOLDER, G. A. GAITANARIS, A. GRAGEROV, M. E. GOTTESMAN, AND S. J. SILVERSTEIN. Inhibition of DnaK autophosphorylation by heat shock proteins and polypeptide substrates. *J. Biol. Chem.* 269: 16643–16647, 1994.

170. PARSELL, D. A., A. S. KOWAL, M. A. SINGER, AND S. LINDQUIST. Protein disaggregation mediated by heat-shock protein Hsp104. *Nature* 372: 475–478, 1994.

171. PELLECCHIA, M., T. SZYPERSKI, D. WALL, C. GEORGOPOULOS, AND K. WUTHRICH. NMR structure of the J-domain and the Gly/Phe-rich region of the *Escherichia coli* DnaJ chaperone. *J. Mol. Biol.* 260: 236–250, 1996.

172. PETERSON, J. R., A. ORA, P. N. VAN, AND A. HELENIUS. Transient, lectin-like association of calreticulin with folding intermediates of cellular and viral glycoproteins. *Mol. Biol. Cell* 6: 1173–1184, 1995.

173. PETIT, M. A., W. BEDALE, J. OSIPIUK, C. LU, M. RAJAGOPALAN, P. McINERNEY, M. F. GOODMAN, AND H. ECHOLS. Sequential folding of UmuC by the Hsp70 and Hsp60 chaperone complexes of *Escherichia coli*. *J. Biol. Chem.* 269: 23824–23829, 1994.

174. PIERPAOLI, E. V., E. SANDMEIER, A. BAICI, H. J. SCHONFELD, S. GISLER, AND P. CHRISTEN. The power stroke of the DnaK/DnaJ/GrpE molecular chaperone system. *J. Mol. Biol.* 269: 757–768, 1997.

175. PRATT, W. B. The role of the Hsp90-based chaperone system in signal transduction by nuclear receptors and receptors signaling via MAP kinase. *Annu. Rev. Pharmacol. Toxicol.* 37: 297–326, 1997.

176. PRIMM, T. P., K. W. WALKER, AND H. F. GILBERT. Facilitated protein aggregation. Effects of calcium on the chaperone and anti-chaperone activity of protein disulfide-isomerase. *J. Biol. Chem.* 271: 33664–33669, 1996.

177. PRIVALOV, P. L. Intermediate states in protein folding. *J. Mol. Biol.* 258: 707–725, 1996.

178. PRODROMOU, C., S. M. ROE, R. O'BRIEN, J. E. LADBURY, P. W. PIPER, AND L. H. PEARL. Identification and structural characterization of the ATP/ADP-binding site in the Hsp90 molecular chaperone. *Cell* 90: 65–75, 1997.

179. PRODROMOU, C., S. M. ROE, P. W. PIPER, AND L. H. PEARL. A molecular clamp in the crystal structure of the N-terminal domain of the yeast Hsp90 chaperone. *Nature Struct. Biol.* 4: 477–482, 1997.

180. PUIG, A., AND H. F. GILBERT. Anti-chaperone behavior of BiP

during the protein disulfide isomerase-catalyzed refolding of reduced denatured lysozyme. *J. Biol. Chem.* 269: 25889–25896, 1994.

181. QIAN, Y. Q., D. PATEL, F. U. HARTL, AND D. J. McCOLL. Nuclear magnetic resonance solution structure of the human Hsp40 (HDJ-1) J-domain. *J. Mol. Biol.* 260: 224–235, 1996.

182. RANSON, N. A., S. G. BURSTON, AND A. R. CLARKE. Binding, encapsulation and ejection: substrate dynamics during a chaperonin-assisted folding reaction. *J. Mol. Biol.* 266: 656–664, 1997.

183. RANSON, N. A., N. J. DUNSTER, S. G. BURSTON, AND A. R. CLARKE. Chaperonins can catalyse the reversal of early aggregation steps when a protein misfolds. *J. Mol. Biol.* 250: 581–586, 1995.

184. REID, B. G., AND G. C. FLYNN. GroEL binds to and unfolds rhodanese posttranslationally. *J. Biol. Chem.* 271: 7212–7217, 1996.

185. REID, K. L., AND A. L. FINK. Physical interactions between members of the DnaK chaperone machinery: characterization of the DnaK. GrpE complex. *Cell Stress Chaperones* 1: 127–137, 1996.

186. RODER, H., AND W. COLON. Kinetic role of early intermediates in protein folding. *Curr. Opin. Struct. Biol.* 7: 15–28, 1997.

187. ROMMELAERE, H., M. VAN TROYS, Y. GAO, R. MELKI, N. J. COWAN, J. VANDEKERCKHOVE, AND C. AMPE. Eukaryotic cytosolic chaperonin contains T-complex polypeptide 1 and seven related subunits. *Proc. Natl. Acad. Sci. USA* 90: 11975–11979, 1993.

188. ROSEMAN, A. M., S. CHEN, H. WHITE, K. BRAIG, AND H. R. SAIBIL. The chaperonin ATPase cycle: mechanism of allosteric switching and movements of substrate-binding domains in GroEL. *Cell* 87: 241–251, 1996.

189. ROWLEY, N., C. PRIP-BUUS, B. WESTERMANN, C. BROWN, E. SCHWARZ, B. BARRELL, AND W. NEUPERT. Mdj1p, a novel chaperone of the DnaJ family, is involved in mitochondrial biogenesis and protein folding. *Cell* 77: 249–259, 1994.

190. RUDDON, R. W., AND E. BEDOWS. Assisted protein folding. *J. Biol. Chem.* 272: 3125–3128, 1997.

191. RUDIGER, S., A. BUCHBERGER, AND B. BUKAU. Interaction of Hsp70 chaperones with substrates. *Nature Struct. Biol.* 4: 342–349, 1997.

192. RUDIGER, S., L. GERMEROTH, J. SCHNEIDER-MERGENER, AND B. BUKAU. Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. *EMBO J.* 16: 1501–1507, 1997.

193. RYAN, M. T., D. J. NAYLOR, P. B. HOJ, M. S. CLARK, AND N. J. HOOGENRAAD. The role of molecular chaperones in mitochondrial protein import and folding. *Int. Rev. Cytol.* 174: 127–193, 1997.

194. RYE, H. S., S. G. BURSTON, W. A. FENTON, J. M. BEECHEM, Z. XU, P. B. SIGLER, AND A. L. HORWICH. Distinct actions of *cis* and *trans* ATP within the double ring of the chaperonin GroEL. *Nature* 388: 792–798, 1997.

195. SADIS, S., AND L. E. HIGHTOWER. Unfolded proteins stimulate molecular chaperone Hsc70 ATPase by accelerating ADP/ATP exchange. *Biochemistry* 31: 9406–9412, 1992.

196. SAIBIL, H. R. What can electron microscopy tell us about chaperoned protein folding? *Folding Design* 1: R45—R49, 1996.

197. SAX, C. M., AND J. PIATIGORSKY. Expression of the alpha-crystallin/small heat-shock protein/molecular. *Adv. Enzymol. Related Areas Mol. Biol.* 69: 155–201, 1994.

198. SCHLENSTEDT, G., S. HARRIS, B. RISSE, R. LILL, AND P. A. SILVER. A yeast DnaJ homologue, Scj1p, can function in the endoplasmic. *J. Cell Biol.* 129: 979–988, 1995.

199. SCHMID, D., A. BAICI, H. GEHRING, AND P. CHRISTEN. Kinetics of molecular chaperone action. *Science* 263: 971–973, 1994.

200. SCHMID, F. X. Catalysis of protein folding by prolyl isomerases In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto. New York: Dekker, 1998, p. 361–389.

201. SCHOLZ, C., G. STOLLER, T. ZARNT, G. FISCHER, AND F. X. SCHMID. Cooperation of enzymatic and chaperone functions of trigger factor in the catalysis of protein folding. *J. Bioenerg. Biomembr.* 29: 35–43, 1997.

202. SCHONFELD, H. J., D. SCHMIDT, H. SCHRODER, AND B. BUKAU. The DnaK chaperone system of *Escherichia coli*: quaternary structures and interactions of the DnaK and GrpE components. *J. Biol. Chem.* 270: 2183–2189, 1995.

203. SCHRODER, H., T. LANGER, F. U. HARTL, AND B. BUKAU. DnaK, DnaJ and GrpE form a cellular chaperone machinery capable of repairing heat-induced protein damage. *EMBO J.* 12: 4137–4144, 1993.

204. SCHUMACHER, R. J., W. J. HANSEN, B. C. FREEMAN, E. ALNEMRI, G. LITWACK, AND D. O. TOFT. Cooperative action of Hsp70, Hsp90, and DnaJ proteins in protein renaturation. *Biochemistry* 35: 14889–14898, 1996.

205. SCHUMACHER, R. J., R. HURST, W. P. SULLIVAN, N. J. McMAHON, D. O. TOFT, AND R. L. MATTS. ATP-dependent chaperoning activity of reticulocyte lysate. *J. Biol. Chem.* 269: 9493–9499, 1994.

206. SECKLER, R. Assembly of oligomers and multisubunit structures. In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto. New York: Dekker, 1998, p. 391–413.

207. SFATOS, C. D., A. M. GUTIN, V. I. ABKEVICH, AND E. I. SHAKHNOVICH. Simulations of chaperone-assisted folding. *Biochemistry* 35: 334–339, 1996.

208. SILOW, M., AND M. OLIVEBERG. High-energy channeling in protein folding. *Biochemistry* 36: 7633–7637, 1997.

209. SONG, J. L., AND C. C. WANG. Chaperone-like activity of protein disulfide-isomerase in the refolding of rhodanese. *Eur. J. Biochem.* 231: 312–316, 1995.

210. SOSNICK, T. R., L. MAYNE, R. HILLER, AND S. W. ENGLANDER. The barriers in protein folding. *Nature Struct. Biol.* 1: 149–156, 1994.

211. SPARRER, H., AND J. BUCHNER. How GroES regulates binding of nonnative protein to GroEL. *J. Biol. Chem.* 272: 14080–14086, 1997.

212. SPARRER, H., H. LILIE, AND J. BUCHNER. Dynamics of the GroEL-protein complex: effects of nucleotides and folding mutants. *J. Mol. Biol.* 258: 74–87, 1996.

213. SPARRER, H., K. RUTKAT, AND J. BUCHNER. Catalysis of protein folding by symmetric chaperone complexes. *Proc. Natl. Acad. Sci. USA* 94: 1096–1100, 1997.

214. SRIRAM, M., J. OSIPIUK, B. FREEMAN, R. MORIMOTO, AND A. JOACHIMIAK. Human Hsp70 molecular chaperone binds two calcium ions within the ATPase domain. *Structure* 5: 403–414, 1997.

215. STERNLICHT, H., G. W. FARR, M. L. STERNLICHT, J. K. DRISCOLL, K. WILLISON, AND M. B. YAFFE. The T-complex polypeptide 1 complex is a chaperonin for tubulin and actin in vivo. *Proc. Natl. Acad. Sci. USA* 90: 9422–9426, 1993.

216. STOLDT, V., F. RADEMACHER, V. KEHREN, J. F. ERNST, D. A. PEARCE, AND F. SHERMAN. Review: the CCT eukaryotic chaperonin subunits of *Saccharomyces cerevisiae* and other yeasts. *Yeast* 12: 523–529, 1996.

217. STRAUS, D., W. WALTER, AND C. A. GROSS. Dnak, DnaJ, and GrpE heat shock proteins negatively regulate heat shock gene expression by controlling the synthesis and stability of sigma 32. *Genes Dev.* 4: 2202–2209, 1990.

218. STRICKLAND, E., B. H. QU, L. MILLEN, AND P. J. THOMAS. The molecular chaperone Hsc70 assists the in vitro folding of the N-terminal nucleotide-binding domain of the cystic fibrosis transmembrane conductance regulator. *J. Biol. Chem.* 272: 25421–25424, 1997.

219. SUH, K., C. A. GABEL, AND J. E. BERGMANN. Identification of a novel mechanism for the removal of glucose residues. *J. Biol. Chem.* 267: 21671–21677, 1992.

220. SZABO, A., R. KORSZUN, F. U. HARTL, AND J. FLANAGAN. A zinc finger-like domain of the molecular chaperone DnaJ is involved in binding to denatured protein substrates. *EMBO J.* 15: 408–417, 1996.

221. SZABO, A., T. LANGER, H. SCHRODER, J. FLANAGAN, B. BUKAU, AND F. U. HARTL. The ATP hydrolysis-dependent reaction cycle of the *Escherichia coli* Hsp70 system DnaK, DnaJ, and GrpE. *Proc. Natl. Acad. Sci. USA* 91: 10345–10349, 1994.

222. SZYPERSKI, T., M. PELLECCHIA, D. WALL, C. GEORGOPOULOS, AND K. WUTHRICH. NMR structure determination of the *Escherichia coli* DnaJ molecular chaperone: secondary structure and backbone fold of the N-terminal region (residues 2–108) containing the highly conserved J domain. *Proc. Natl. Acad. Sci. USA* 91: 11343–11347, 1994.

223. TAKAYAMA, S., D. N. BIMSTON, S. MATSUZAWA, B. C. FREEMAN, C. AIME-SEMPE, Z. XIE, R. I. MORIMOTO, AND J. C. REED. BAG-1 modulates the chaperone activity of Hsp70/Hsc70. *EMBO J.* 16: 4887–4896, 1997.

224. TAKEDA, S., AND D. B. McKAY. Kinetics of peptide binding to the

bovine 70 kDa heat shock cognate protein, a molecular chaperone. *Biochemistry* 35: 4636–4644, 1996.

225. TAKENAKA, I. M., S. M. LEUNG, S. J. McANDREW, J. P. BROWN, AND L. E. HIGHTOWER. Hsc70-binding peptides selected from a phage display peptide library that resemble organellar targeting sequences. *J. Biol. Chem.* 270: 19839–19844, 1995.

226. TAVARIA, M., T. GABRIELE, I. KOLA, AND R. L. ANDERSON. A hitchhiker's guide to the human Hsp70 family. *Cell Stress Chaperones* 1: 23–28, 1996.

227. THEYSSEN, H., H. P. SCHUSTER, L. PACKSCHIES, B. BUKAU, AND J. REINSTEIN. The second step of ATP binding to DnaK induces peptide release. *J. Mol. Biol.* 263: 657–670, 1996.

228. THOMAS, J. G., A. AYLING, AND F. BANEYX. Molecular chaperones, folding catalysts, and the recovery of active recombinant proteins from *E. coli* to fold or to refold. *Appl. Biochem. Biotechnol.* 66: 197–238, 1997.

229. THOMAS, J. G., AND F. BANEYX. Protein folding in the cytoplasm of *Escherichia coli*: requirements for the DnaK-DnaJ-GrpE and GroEL-GroES molecular chaperone machines. *Mol. Microbiol.* 21: 1185–1196, 1996.

230. TIAN, G., Y. HUANG, H. ROMMELAERE, J. VANDEKERCKHOVE, C. AMPE, AND N. J. COWAN. Pathway leading to correctly folded beta-tubulin. *Cell* 86: 287–296, 1996.

231. TIAN, G., S. A. LEWIS, B. FEIERBACH, T. STEARNS, H. ROMMELAERE, C. AMPE, AND N. J. COWAN. Tubulin subunits exist in an activated conformational state generated and maintained by protein cofactors. *J. Cell Biol.* 138: 821–832, 1997.

232. TODD, M. J., G. H. LORIMER, AND D. THIRUMALAI. Chaperonin-facilitated protein folding: optimization of rate and yield by an iterative annealing mechanism. *Proc. Natl. Acad. Sci. USA* 93: 4030–4035, 1996.

233. TODD, M. J., P. V. VIITANEN, AND G. H. LORIMER. Hydrolysis of adenosine 5′-triphosphate by *Escherichia coli* GroEL: effects of GroES and potassium ion. *Biochemistry* 32: 8560–8567, 1993.

234. TODD, M. J., P. V. VIITANEN, AND G. H. LORIMER. Dynamics of the chaperonin ATPase cycle: implications for facilitated protein folding. *Science* 265: 659–666, 1994.

235. TOKATLIDIS, K., B. FRIGUET, D. DEVILLE-BONNE, F. BALEUX, A. N. FEDOROV, A. NAVON, L. DJAVADI-OHANIANCE, AND M. E. GOLDBERG. Nascent chains: folding and chaperone interaction during elongation on ribosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 348: 89–95, 1995.

236. TOPPING, T. B., AND L. L. RANDALL. Chaperone SecB from *Escherichia coli* mediates kinetic partitioning via a dynamic equilibrium with its ligands. *J. Biol. Chem.* 272: 19314–19318, 1997.

237. TOROK, Z., L. VIGH, AND P. GOLOUBINOFF. Fluorescence detection of symmetric GroEL$_{14}$(GroES7)$_2$ heterooligomers involved in protein release during the chaperonin cycle. *J. Biol. Chem.* 271: 16180–16186, 1996.

238. TSAI, J., AND M. G. DOUGLAS. A conserved HPD sequence of the J-domain is necessary for YDJ1 stimulation of Hsp70 ATPase activity at a site distinct from substrate binding. *J. Biol. Chem.* 271: 9347–9354, 1996.

239. UNGEWICKELL, E., H. UNGEWICKELL, S. E. HOLSTEIN, R. LINDNER, K. PRASAD, W. BAROUCH, B. MARTIN, L. E. GREENE, AND E. EISENBERG. Role of auxilin in uncoating clathrin-coated vesicles. *Nature* 378: 632–635, 1995.

240. VALENT, Q. A., D. A. KENDALL, S. HIGH, R. KUSTERS, B. OUDEGA, AND J. LUIRINK. Early events in preprotein recognition in *E. coli*: interaction of SRP and trigger factor with nascent polypeptides. *EMBO J.* 14: 5494–5505, 1995.

241. VASSILAKOS, A., M. F. COHEN-DOYLE, P. A. PETERSON, M. R. JACKSON, AND D. B. WILLIAMS. The molecular chaperone calnexin facilitates folding and assembly of class I histocompatibility molecules. *EMBO J.* 15: 1495–1506, 1996.

242. VICKERY, L. E., J. J. SILBERG, AND D. T. TA. Hsc66 and Hsc20, a new heat shock cognate molecular chaperone system from *Escherichia coli*. *Protein Sci.* 6: 1047–1056, 1997.

243. VIITANEN P. V., T. H. LUBBEN, J. REED, P. GOLOUBINOFF, D. P. O'KEEFE, AND G. H. LORIMER. Chaperonin-facilitated refolding of ribulosebisphosphate carboxylase and ATP hydrolysis by chaperonin 60 (GroEL) are K$^+$ dependent. *Biochemistry* 29: 5665–5671, 1990.

244. VYSOKANOV, A. V. Synthesis of chloramphenicol acetyltransferase in a coupled transcription-translation in vitro system lacking the chaperones DnaK and DnaJ. *FEBS Lett.* 375: 211–214, 1995.

245. WALKER, K. W., AND H. F. GILBERT. Protein disulfide isomerase In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto. New York: Dekker, 1998, p. 331–359.

246. WALL, D., M. ZYLICZ, AND C. GEORGOPOULOS. The conserved G/F motif of the DnaJ chaperone is necessary for the activation of the substrate binding properties of the DnaK chaperone. *J. Biol. Chem.* 270: 2139–2144, 1995.

247. WALTER, S., G. H. LORIMER, AND F. X. SCHMID. A thermodynamic coupling mechanism for GroEL-mediated unfolding. *Proc. Natl. Acad. Sci. USA* 93: 9425–9430, 1996.

248. WANG, J., J. ONUCHIC, AND P. WOLYNES. Statistics of kinetic pathways on biased rough energy landscapes with applications to protein folding. *Phys. Rev. Lett.* 76: 4861–4864, 1996.

249. WAWRZYNOW, A., B. BANECKI, D. WALL, K. LIBEREK, C. GEORGOPOULOS, AND M. ZYLICZ. ATP hydrolysis is required for the DnaJ-dependent activation of DnaK chaperone for binding to both native and denatured protein substrates. *J. Biol. Chem.* 270: 19307–19311, 1995.

250. WEI, J., AND L. M. HENDERSHOT. Protein folding and assembly in the endoplasmic reticulum. *EXS* 77: 41–55, 1996.

251. WEISSMAN, J. S., C. M. HOHL, O. KOVALENKO, Y. KASHI, S. CHEN, K. BRAIG, H. R. SAIBIL, W. A. FENTON, AND A. L. HORWICH. Mechanism of GroEL action: productive release of polypeptide from a sequestered position under GroES. *Cell* 83: 577–587, 1995.

252. WELCH, W. J., AND C. R. BROWN. Influence of molecular and chemical chaperones on protein folding. *Cell Stress Chaperones* 1: 109–115, 1996.

253. WETZEL, R. Principles of protein stability. Part 2: enhanced folding and stabilization of proteins by suppression of aggregation in vitro and in vivo. In: *Protein Engineering: A Practical Approach*, edited by A. R. Rees and M. J. E. Sternberg. New York: Oxford Univ. Press, 1992, p. 191–221.

254. WETZEL, R. Mutations and off-pathway aggregation of proteins. *Trends Biotechnol.* 12: 193–198, 1994.

255. WETZEL, R. For protein misassembly, it's the "I" decade. *Cell* 86: 699–702, 1996.

256. WHITE, Z. W., K. E. FISHER, AND E. EISENSTEIN. A monomeric variant of GroEL binds nucleotides but is inactive as a molecular chaperone. *J. Biol. Chem.* 270: 20404–20409, 1995.

257. WICKNER, S., S. GOTTESMAN, D. SKOWYRA, J. HOSKINS, K. McKENNEY, AND M. R. MAURIZI. A molecular chaperone, ClpA, functions like DnaK and DnaJ. *Proc. Natl. Acad. Sci. USA* 91: 12218–12222, 1994.

258. WILD, J., P. ROSSMEISSL, W. A. WALTER, AND C. A. GROSS. Involvement of the DnaK-DnaJ-GrpE chaperone team in protein secretion in *Escherichia coli*. *J. Bacteriol.* 178: 3608–3613, 1996.

259. WILLIAMS, D. B. The Merck Frosst Award Lecture 1994/La conference Merck Frosst 1994. Calnexin: a molecular chaperone with a taste for carbohydrate. *Biochem. Cell. Biol.* 73: 123–132, 1995.

260. WISNIEWSKI, T., A. GOLABEK, E. MATSUBARA, J. GHISO, AND B. FRANGIONE. Apolipoprotein E: binding to soluble Alzheimer's beta-amyloid. *Biochem. Biophys. Res. Commun.* 192: 359–365, 1993.

261. WIUFF, C., AND G. HOUEN. Cation-dependent interactions of calreticulin with denatured and native proteins. *Acta Chem. Scand.* 50: 788–795, 1996.

262. WU, B., A. WAWRZYNOW, M. ZYLICZ, AND C. GEORGOPOULOS. Structure-function analysis of the *Escherichia coli* GrpE heat shock protein. *EMBO J.* 15: 4806–4816, 1996.

263. XU, Z. H., A. L. HORWICH, AND P. B. SIGLER. The crystal structure of the asymmetric GroEL-GroES-(ADP)$_7$ chaperonin complex. *Nature* 388: 741–750, 1997.

264. YAFFE, M. B., G. W. FARR, D. MIKLOS, A. L. HORWICH, M. L. STERNLICHT, AND H. STERNLICHT. TCP1 complex is a molecular chaperone in tubulin biogenesis. *Nature* 358: 245–248, 1992.

265. YAHARA, I. Structure and function of the 90-kDa stress protein Hsp90. In: *Molecular Chaperones in the Life Cycle of Proteins*, edited by A. L. Fink and Y. Goto. New York: Dekker, 1998, p. 183–192.

266. YIFRACH, O., AND A. HOROVITZ. Nested cooperativity in the ATPase activity of the oligomeric chaperonin GroEL. *Biochemistry* 34: 5303–5308, 1995.

267. YON, J. M. Protein folding: concepts and perspectives. *Cell. Mol. Life Sci.* 53: 557–567, 1997.

268. ZAHN, R., S. PERRETT, AND A. R. FERSHT. Conformational states bound by the molecular chaperones GroEL and secB: a hidden unfolding (annealing) activity. *J. Mol. Biol.* 261: 43–61, 1996.

269. ZAHN, R., S. PERRETT, G. STENBERG, AND A. R. FERSHT. Catalysis of amide proton exchange by the molecular chaperones GroEL and SecB. *Science* 271: 642–645, 1996.

270. ZARNT, T., T. TRADLER, G. STOLLER, C. SCHOLZ, F. X. SCHMID, AND G. FISCHER. Modular structure of the trigger factor required for high activity in protein folding. *J. Mol. Biol.* 271: 827–837, 1997.

271. ZEINER, M., M. GEBAUER, AND U. GEHRING. Mammalian protein RAP46: an interaction partner and modulator of 70 kDa heat shock proteins. *EMBO J.* 16: 5483–5490, 1997.

272. ZHU, X., X. ZHAO, W. F. BURKHOLDER, A. GRAGEROV, C. M. OGATA, M. E. GOTTESMAN, AND W. A. HENDRICKSON. Structural analysis of substrate binding by the molecular chaperone DnaK. *Science* 272: 1606–1614, 1996.

273. ZIEGELHOFFER, T., P. LOPEZ-BUESA, AND E. A. CRAIG. The dissociation of ATP from hsp70 of *Saccharomyces cerevisiae* is stimulated by both Ydj1p and peptide substrates. *J. Biol. Chem.* 270: 10412–10419, 1995.

274. ZYLICZ, M., T. YAMAMOTO, N. McKITTRICK, S. SELL, AND C. GEORGOPOULOS. Purification and properties of the DnaJ replication protein of *Escherichia coli*. *J. Biol. Chem.* 260: 7591–7598, 1985.

Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 8945–8947, August 1997
Applied Biological Sciences

Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR

DEVAL A. LASHKARI*†, JOHN H. MCCUSKER‡, AND RONALD W. DAVIS*§

*Departments of Genetics and Biochemistry, Beckman Center, Stanford University, Stanford, CA 94305; and ‡Department of Microbiology, 3020 Duke University Medical Center, Durham, NC 27710

Contributed by Ronald W. Davis, May 20, 1997

ABSTRACT **The recent ability to sequence whole genomes allows ready access to all genetic material. The approaches outlined here allow automated analysis of sequence for the synthesis of optimal primers in an automated multiplex oligonucleotide synthesizer (AMOS). The efficiency is such that all ORFs for an organism can be amplified by PCR. The resulting amplicons can be used directly in the construction of DNA arrays or can be cloned for a large variety of functional analyses. These tools allow a replacement of single-gene analysis with a highly efficient whole-genome analysis.**

The genome sequencing projects have generated and will continue to generate enormous amounts of sequence data. The genomes of *Saccharomyces cerevisiae*, *Escherichia coli*, *Haemophilus influenzae* (1), *Mycoplasma genitalium* (2), and *Methanococcus jannaschii* (3) have been completely sequenced. Other model organisms have had substantial portions of their genomes sequenced as well, including the nematode *Caenorhabditis elegans* (4) and the small flowering plant *Arabidopsis thaliana* (5). This massive and increasing amount of sequence information allows the development of novel experimental approaches to identify gene function.

One standard use of genome sequence data is to attempt to identify the functions of predicted open reading frames (ORFs) within the genome by comparison to genes of known function. Such a comparative analysis of all ORFs to existing sequence data is fast, simple, and requires no experimentation and is therefore a reasonable first step. While finding sequence homologies/motifs is not a substitute for experimentation, noting the presence of sequence homology and/or sequence motifs can be a useful first step in finding interesting genes, in designing experiments and, in some cases, predicting function. However, this type of analysis is frequently uninformative. For example, over one-half of new ORFs in *S. cerevisiae* have no known function (6). If this is the case in a well studied organism such as yeast, the problem will be even worse in organisms that are less well studied or less manipulable. A large, experimentally determined gene function database would make homology/motif searches much more useful.

Experimental' analysis must be performed to thoroughly understand the biological function of a gene product. Scaling up from classical "cottage industry" one-gene-oriented approaches to whole-genome analysis would be very expensive and laborious. It is clear that novel strategies are necessary to efficiently pursue the next phase of the genome projects—whole-genome experimental analysis to explore gene expression, gene product function, and other genome functions. Model organisms, such as *S. cerevisiae*, will be extremely important in the development of novel whole-genome analysis techniques and, subsequently, in improving our understanding of other more complex and less manipulable organisms.

The genome sequence can be systematically used as a tool to understand ORFs, gene product function, and other genome regions. Toward this end, a directed strategy has been developed for exploiting sequence information as a means of providing information about biological function (Fig. 1). Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons—they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay (7). As a pilot study, synthetic primers were made on the 96-well automated multiplex oligonucleotide synthesizer (AMOS) instrument (8) (Fig. 2). These oligonucleotides were used to amplify each ORF on yeast chromosome V. The current version of this instrument can synthesize three plates of 96 oligonucleotides each (25 bases) in an 8-hr day. The amplification of the entire set of PCR products was then analyzed by gel electrophoresis (Fig. 3). Successful amplification of the proper length product on the first attempt was 95%. This project demonstrates that one can go directly from sequence information to biological analysis in a truly automated, totally directed manner.

These amplicons can be incorporated directly in arrays or the amplicons can be cloned. If the amplicons are to be cloned, novel sequences can be incorporated at the 5′ end of the oligonucleotide to facilitate cloning. One potential problem with cloning PCR products is that the cloned amplicons may contain sequence alterations that diminish their utility. One option would be to resequence each individual amplicon. However, this is expensive, inefficient, and time consuming. A faster, more cost-effective, and more accurate approach is to apply comparative sequencing by denaturing HPLC (9). This method is capable of detecting a single base change in a 2-kb heteroduplex. Longer amplicons can be analyzed by use of appropriate restriction fragments. If any change is detected in a clone, an alternate clone of the same region can be analyzed. Modifying the system to allow high throughput analysis by denaturing HPLC is also relatively simple and straightforward.

If amplicons are used directly on arrays without cloning, it is important to note that, even if single PCR product bands are observed on gels, the PCR products will be contaminated with various amounts of other sequences. This contamination has the potential to affect the results in, for example, expression

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at http://www.pnas.org.

†Present address: Synteni, Inc., 6519 Dumbarton Circle, Fremont, CA 94555.

§To whom reprint requests should be addressed at: Department of Biochemistry, Beckman Center, B400, Stanford University, Stanford, CA 94305-5307. e-mail: gilbert@cmgm.stanford.edu.

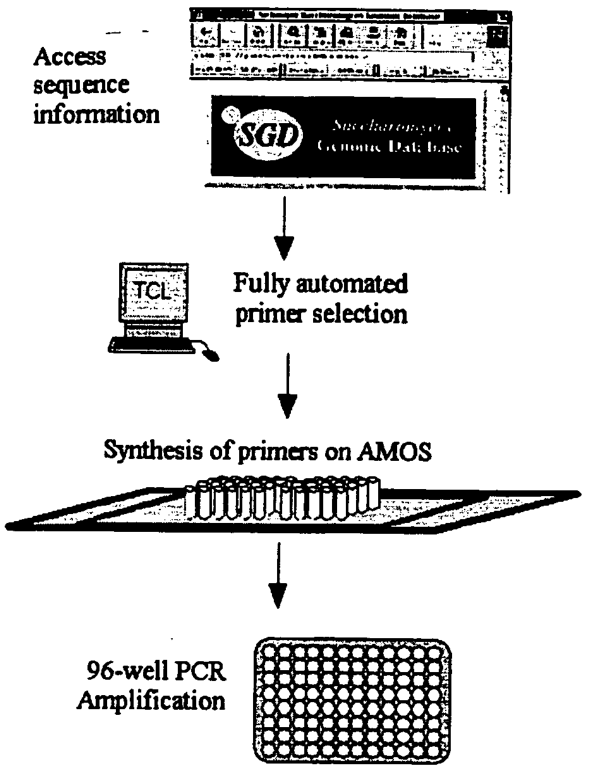

Access sequence information
SGD
Saccharomyces Genome Data base
TCL
Fully automated primer selection
Synthesis of primers on AMOS
96-well PCR Amplification

FIG. 1. Overview of systematic method for isolating individual genes. Sequence information is obtained automatically from sequence databases. The data are input into primer selection software specifically designed to target ORFs as designated by database annotations. The output file containing the primer information is directly read by a high-throughput oligonucleotide synthesizer, which makes the oligonucleotides in 96-well plates (AMOS, automated multiplex oligonucleotide synthesizer). The forward and reverse primers are synthesized in the same location on separate plates to facilitate the downstream handling of primers. The amplicons are generated by PCR in 96-well plates as well.

analysis. On the other hand, direct use of the amplicons is much less labor intensive and greatly decreases the occurrence of mistakes in clone identification, a ubiquitous problem associated with large clone set archiving and retrieving.

Any large-scale effort to capture each ORF within a genome must rely on automation if cost is to be minimized while efficiency is maximized. Toward that end, primers targeting ORFs were designed automatically using simple new scripts and existing primer selection software. These script-selected primer sequences were directly read by the high-throughput synthesizer and the forward and reverse primers were synthesized in separate plates in corresponding wells to facilitate automated pipetting and PCR amplifications. Each of the resulting PCR products, generated with minimum labor, contains a known, unique ORF.

Large-scale genome analysis projects are dependent on newly emerging technologies to make the studies practical and economically feasible. For example, the cost of the primers, a significant issue in the past, has been reduced dramatically to make feasible this and other projects that require tens of thousands of oligonucleotides. Other methods of high-throughput analysis are also vital to the success of functional analysis projects, such as microarraying and oligonucleotide chip methods (10–14).

Changes in attitude are also required. One of the major costs of commercial oligonucleotides is extensive quality control such that virtually 100% of the supplied oligonucleotides are successfully synthesized and work for their intended purpose. Considerable cost reduction can be obtained by simply decreasing the expected successful synthesis rate to 95–97%. One can then achieve faster and cheaper whole genome coverage by simply adding a single quality control at the end of the experiment and batching the failures for resynthesis.

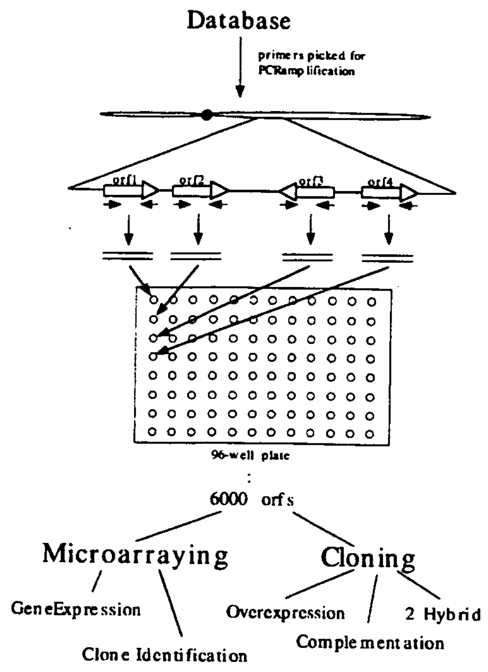


FIG. 2. Overall approach for using database of a genome to direct biological analysis. The synthesis of the 6,000 ORFs (orfs) for each gene of *S. cerevisiae* can be used in many applications utilizing both cloning and microarraying technology.

The directed nature of the amplicon approach is of clear advantage. The sequence of each ORF is analyzed automatically, and unique specific primers are made to target each ORF. Thus, there is relatively little time or labor involved—for example, no random cloning and subsequent screening is required because each product is known. In the test system, primers for 240 ORFs from chromosome V were systematically synthesized, beginning from the left arm and continuing through to the right arm. At no point was there any manual analysis of sequence information to generate the collection. In many ways, now that the sequence is known, there is no need for the researcher to examine it.

These amplicons can be arrayed and expression analysis can be done on all arrayed ORFs with a single hybridization (10). Those ORFs that display significant differential expression patterns under a given selection are easily identified without the laborious task of searching for and then sequencing a clone. Once scaled up, the procedure provides even greater returns on effort, because a single hybridization will ultimately provide a "snapshot" of the expression of all genes in the yeast genome. Thus, the limiting factor in whole genome analysis will not be the analysis process itself, but will instead be the ability of researchers to design and carry out experimental selections.

Current expression and genetic analysis technologies are geared toward the analysis of single genes and are ill suited to analyze numerous genes under many conditions. Additional difficulties with current technologies include: the effort and expense required to analyze expression and make mutants, the potential duplication of effort if done by different laboratories, and the possibility of conflicting results obtained from different laboratories. In contrast, whole genome analysis not only is more efficient, it also provides data of much higher quality; all genes are assayed and compared in parallel under exactly

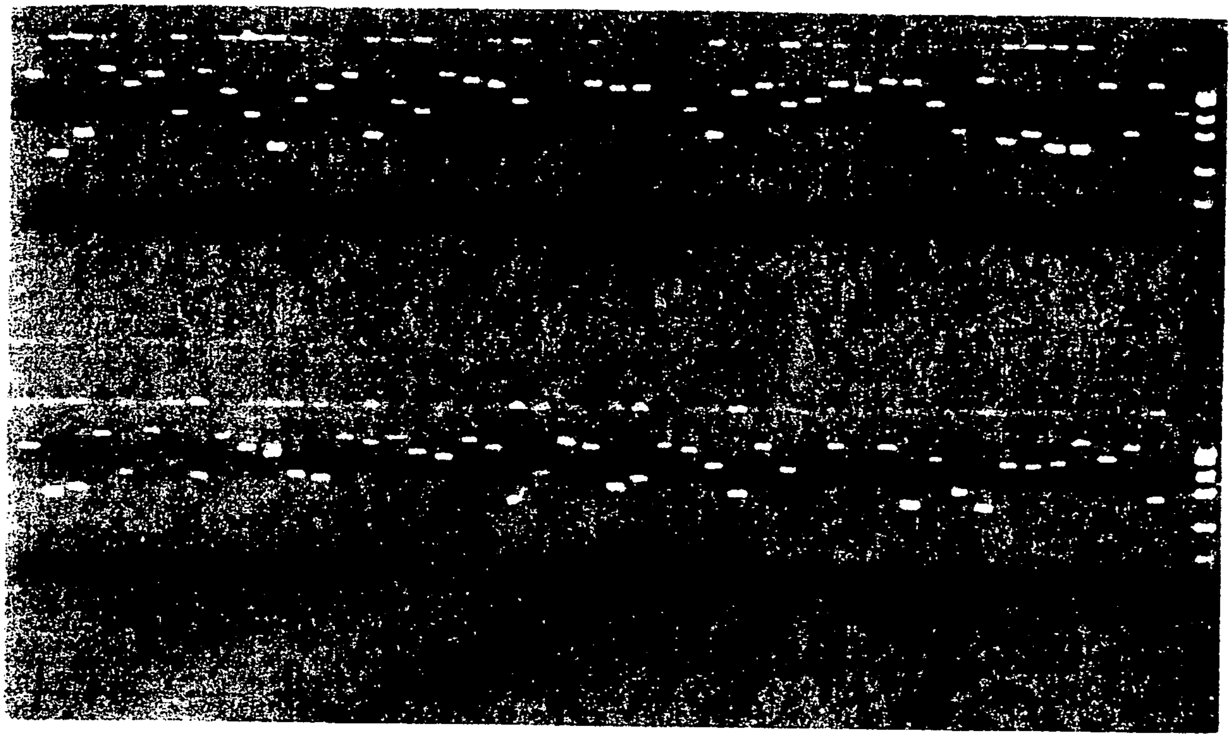


FIG. 3. Gel image of amplifications. Using the method described in Fig. 1, amplicons were generated for ORFs of *S. cerevisiae* chromosome V. One plate of 96 amplification reactions is shown.

the same conditions. In addition, amplicons have many applications beyond gene expression. For example, one recent approach is to incorporate a unique DNA sequence tag, synthesized as part of each gene specific primer, during amplification. The tags or molecular bar codes, when reintroduced into the organism as a gene deletion or as a gene clone, can be used much more efficiently than individual mutations or clones because pools of tagged mutants or transformants can be analyzed in parallel. This parallel analysis is possible because the tags are readily and quantitatively amplified even in complex mixtures of tags (13).

These ORF genome arrays and oligonucleotide tagged libraries can be used for many applications. Any conventional selection applied to a library that gives discrete or multiple products can use these technologies for a simple direct readout. These include screens and selections for mutant complementation, overexpression suppression (15, 16), second-site suppressors, synthetic lethality, drug target overexpression (17), two-hybrid screens (18), genome mismatch scanning (19), or recombination mapping.

The genome projects have provided researchers with a vast amount of information. These data must be used efficiently and systematically to gain a truly comprehensive understanding of gene function and, more broadly, of the entire genome which can then be applied to other organisms. Such global approaches are essential if we are to gain an understanding of the living cell. This understanding should come from the viewpoint of the integration of complex regulatory networks, the individual roles and interactions of thousands of functional gene products, and the effect of environmental changes on both gene regulatory networks and the roles of all gene products. The time has come to switch from the analysis of a single gene to the analysis of the whole genome.

Support was provided by National Institutes of Health Grants R37H60198 and P01H600205.

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* **269,** 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* **270,** 397–403.
3. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., *et al.* (1996) *Science* **273,** 1058–1073.
4. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. & Waterston, R. (1992) *Nature (London)* **356,** 37–41.
5. Newman, T., de Bruijn, F. J., Green, P., Keegstra, K., Kende, H., *et al.* (1994) *Plant Physiol.* **106,** 1241–1255.
6. Oliver, S. (1996) *Nature (London)* **379,** 597–600.
7. Lashkari, D. A. (1996) Ph.D. dissertation (Stanford Univ., Stanford, CA).
8. Lashkari, D. A., Hunicke-Smith, S. P., Norgren, R. M., Davis, R. W. & Brennan, T. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 7912–7915.
9. Oefner, P. J. & Underhill, P. A. (1995) *Am. J. Hum. Genet.* **57,** A266.
10. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270,** 467–470.
11. Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. (1991) *Science* **251,** 767–773.
12. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. & Fodor, S. P. (1996) *Science* **274,** 610–614.
13. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. (1996) *Nat. Genet.* **14,** 450–456.
14. Smith, V., Chou, K., Lashkari, D., Botstein, D. & Brown, P. O. (1996) *Science* **274,** 2069–2074.
15. Magdolen, V., Drubin, D. G., Mages, G. & Bandlow, W. (1993) *FEBS Lett.* **316,** 41–47.
16. Ramer, S. W., Elledge, S. J. & Davis, R. W. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 11589–11593.
17. Rine, J., Hansen, W., Hardeman, E. & Davis, R. W. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 6750–6754.
18. Fields, S. & Song, O. (1989) *Nature (London)* **340,** 245–246.
19. Nelson, S. F., McCusker, J. H., Sander, M. A., Kee, Y., Modrich, P. & Brown, P. O. (1994) *Nat. Genet.* **4,** 11–18.

Reference 2 of 9
with Response dated 4/23/04
In USSN: 10/049,742

MOLECULAR CARCINOGENESIS 24:153–159 (1999)

IN PERSPECTIVE
Claudio J. Conti, Editor

Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,[1] Michael Bittner,[2] Jeffrey Trent,[2] J. Carl Barrett,[1] and Cynthia A. Afshari[1]

[1] *Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina*
[2] *Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland*

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog. 24:153–159, 1999.* © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae,* was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cervisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNAse protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10–12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

MICROARRAY DEVELOPMENT AND APPLICATIONS

cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3′ regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluors. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease–related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cervisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22–24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25–27].

Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28–30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnostics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only 4n cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)+ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and BRCA1 [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

THE USE OF MICROARRAYS IN TOXICOLOGY

Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on relative fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

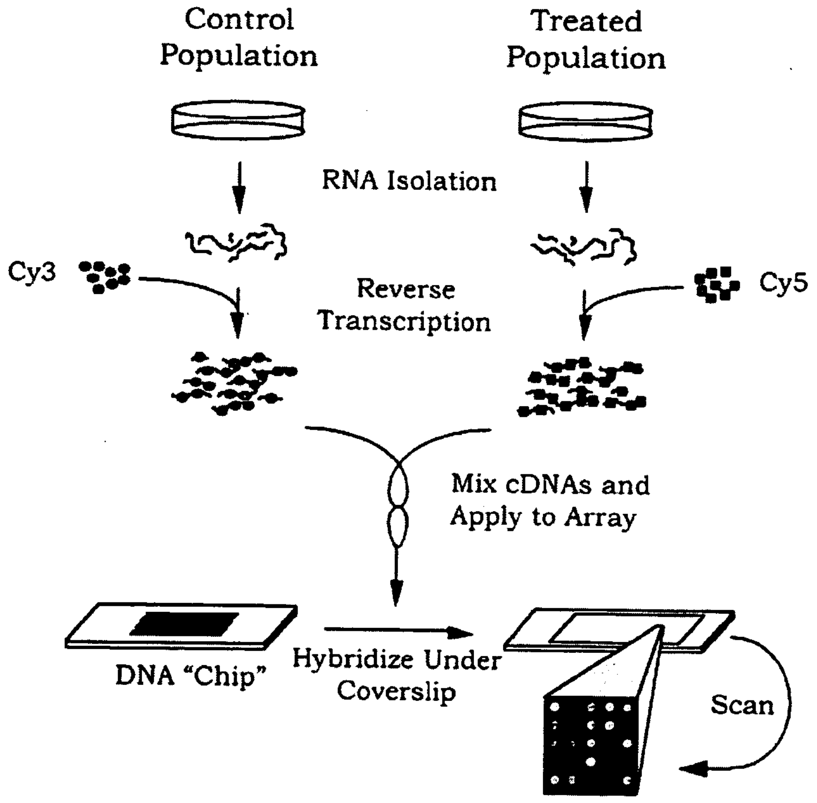


Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illustrative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

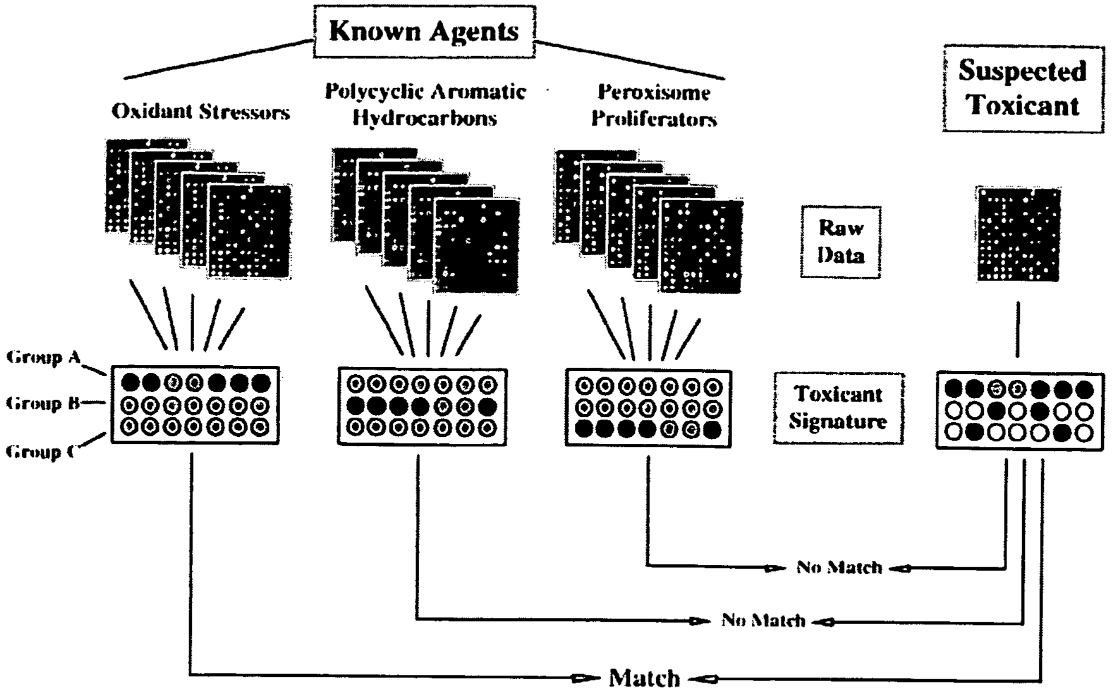


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive in vivo test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia,* and *Arabadopsis* could also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker,Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

REFERENCES

1. http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html
2. http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496–512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. Science 1996;274:546, 563–567.
5. http://www.perkin-elmer.com/press/prc5448.html
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992;257:967–971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res 1996;6:492–503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. Gene 1995;156:207–213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270:484–487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. Science 1995;270:467–470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457–460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol 1997;15:1359–1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. Nat Biotechnol 1998;16:27–31.
14. http://www.synteni.com
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 1996;6:639–645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Biomedical Optics 1997;2:364–374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 1998;58:5009–5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 1996; 93:10614–10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997;94:13057–13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci USA 1997;94:2150–2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680–686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. Genomics 1996;37:29–40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the E. coli genome. Genomics 1996;37:77–86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. Biotechniques 1994;17:328–329, 332–336.
25. http://www.resgen.com/
26. http://www.genomesystems.com/
27. http://www.clontech.com/
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA 1994;91:5022–5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251:767–773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. Proc Natl Acad Sci USA 1996;93:13555–13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 1995;19:442–447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675–1680.
34. http://www.mdyn.com/
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. Genomics 1996;33:445–456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274:610–614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. Nat Genet 1998;18:155–158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum Mutat 1996;7:244–255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. Nat Genet 1996;14:441–447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat Med 1996;2:753–759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077–1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. Science 1998;281:1194–1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. Environ Health Perspect 1990;86:313–321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. Nat Genet 1998;20:19–23.
45. http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 1996;382:722–725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. J Natl Cancer Inst 1993;85:1159–1164.
48. http://www.niehs.nih.gov/envgenom/home.html

ELSEVIER

Toxicology Letters 112–113 (2000) 467–471

Toxicology Letters
www.elsevier.com/locate/toxlet

Expression profiling in toxicology — potentials and limitations

Sandra Steiner *, N. Leigh Anderson

Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA

Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers.

Keywords: Proteomics; Genomics; Toxicology

1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

* Corresponding author. Tel.: +1-301-4245989; fax: +1-301-7624892.

E-mail address: steiner@lsbc.com (S. Steiner)

0378-4274/00/$ - see front matter © 2000 Elsevier Science Ireland Ltd. All rights reserved.
PII: S0378-4274(99)00236-2

2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

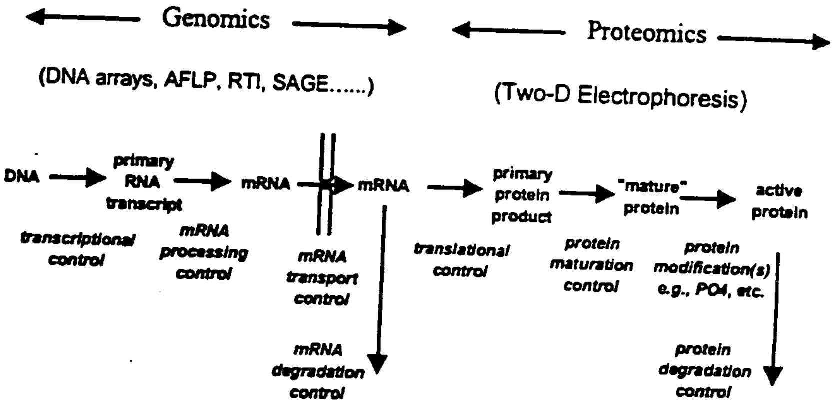


Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.

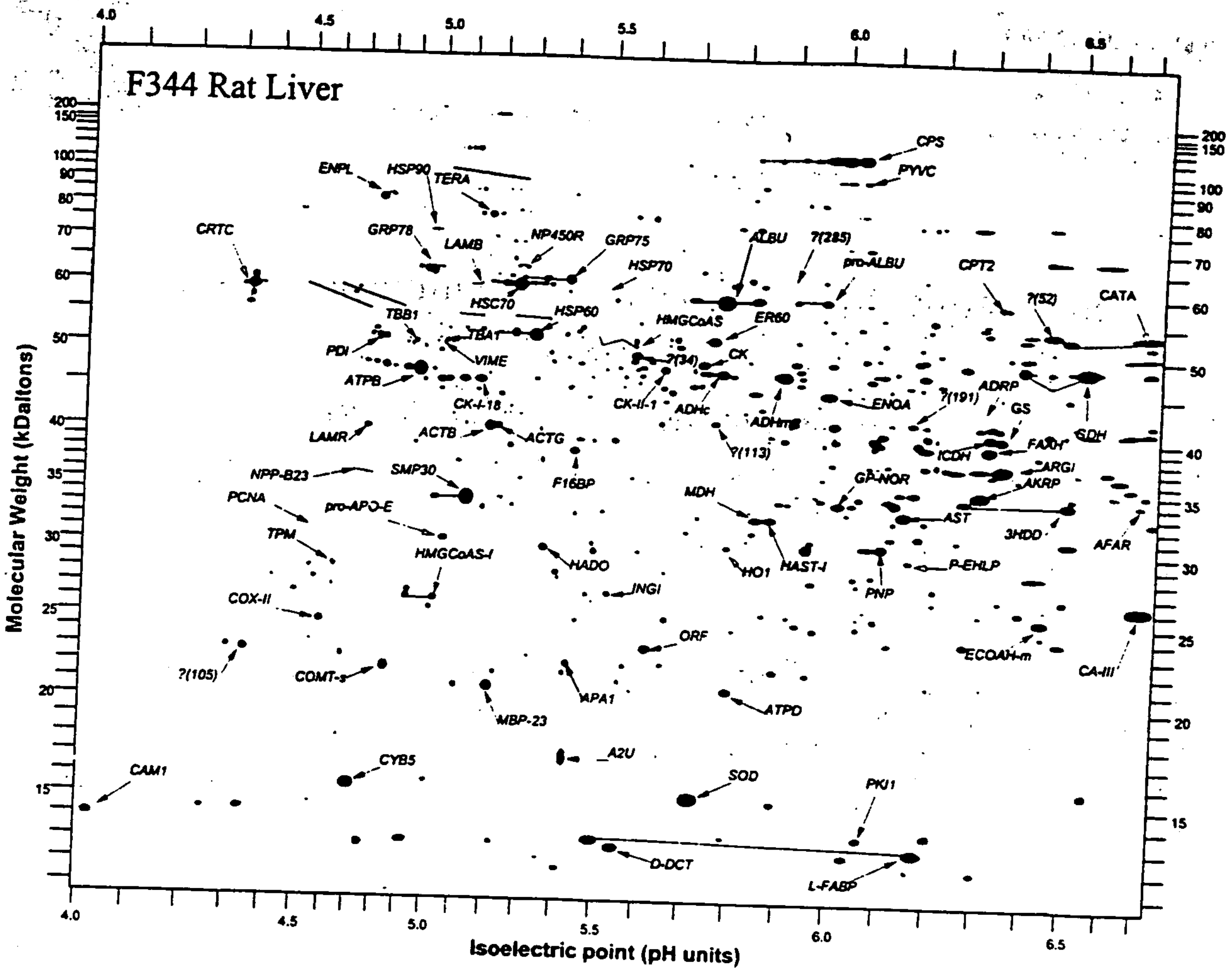


Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a drug candidate.

5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target amplification. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug development, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trails.

References

Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. Electrophoresis 19, 1998–2003.

Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18, 533–537.

Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. Electrophoresis 12, 907–930.

Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. J. Cell. Biochem. Suppl. 22, 108–116.

Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. Toxicol. Appl. Pharmacol. 137, 75–89.

Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. Life Sci. 63, 2243–2250.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. Science 274, 610–614.

Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. Electrophoresis 19, 355–363.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. Science 251, 767–773.

Mann, M., Hojrup, P., Roepsdorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22, 338–345.

Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. Electrophoresis 14, 157–161.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expresssion patterns with a complementary DNA microarray. Science 251, 467–470.

Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 6, 639–645.

Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Raynackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. Biochem. Biophys. Res. Commun. 218, 777–782.

Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. Biochem. Pharmacol. 51, 253–258.

Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. Curr. Biol. 6, 1543–1544.

Application of DNA Arrays to Toxicology

John C. Rockett and David J. Dix

Reproductive Toxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

DNA array technology makes it possible to rapidly genotype individuals or quantify the expression of thousands of genes on a single filter or glass slide, and holds enormous potential in toxicologic applications. This potential led to a U.S. Environmental Protection Agency-sponsored workshop titled "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. In addition to providing state-of-the-art information on the application of DNA or gene microarrays, the workshop catalyzed the formation of several collaborations, committees, and user's groups throughout the Research Triangle Park area and beyond. Potential application of microarrays to toxicologic research and risk assessment include genome-wide expression analyses to identify gene-expression networks and toxicant-specific signatures that can be used to define mode of action, for exposure assessment, and for environmental monitoring. Arrays may also prove useful for monitoring genetic variability and its relationship to toxicant susceptibility in human populations. *Key words:* DNA arrays, gene arrays, microarrays, toxicology. ***Environ Health Perspect*** 107:681–685 (1999). [Online 6 July 1999]
http://ehpnet1.niehs.nih.gov/docs/1999/107p681-685rockett/abstract.html

Decoding the genetic blueprint is a dream that offers manifold returns in terms of understanding how organisms develop and function in an often hostile environment. With the rapid advances in molecular biology over the last 30 years, the dream has come a step closer to reality. Molecular biologists now have the ability to elucidate the composition of any genome. Indeed, almost 20 genomes have already been sequenced and more than 60 are currently under way. Foremost among these is the Human Genome Mapping Project. However, the genomes of a number of commonly used laboratory species are also under intensive investigation, including yeast, *Arabidopsis*, maize, rice, zebra fish, mouse, rat, and dog. It is widely expected that the completion of such programs will facilitate the development of many powerful new techniques and approaches to diagnosing and treating genetically and environmentally induced diseases which afflict mankind. However, the vast amount of data being generated by genome mapping will require new high-throughput technologies to investigate the function of the millions of new genes that are being reported. Among the most widely heralded of the new functional genomics technologies are DNA arrays, which represent perhaps the most anticipated new molecular biology technique since polymerase chain reaction (PCR).

Arrays enable the study of literally thousands of genes in a single experiment. The potential importance of arrays is enormous and has been highlighted by the recent publication of an entire *Nature Genetics* supplement dedicated to the technology (*1*). Despite this huge surge of interest, DNA arrays are still little used and largely unproven, as demonstrated by the high ratio of review and press articles to actual data papers. Even so, the potential they offer has driven venture capitalists into a frenzy of investment and many new companies are springing up to claim a share of this rapidly developing market.

The U.S. Environmental Protection Agency (EPA) is interested in applying DNA array technology to ongoing toxicologic studies. To learn more about the current state of the technology, the Reproductive Toxicology Division (RTD) of the National Health and Environmental Effects Research Laboratory (NHEERL; Research Triangle Park, NC) hosted a workshop on "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. The workshop was organized by David Dix, Robert Kavlock, and John Rockett of the RTD/NHEERL. Twenty-two intramural and extramural scientists from government, academia, and industry shared information, data, and opinions on the current and future applications for this exciting new technology. The workshop had more than 150 attendees, including researchers, students, and administrators from the EPA, the National Institute of Environmental Health Sciences (NIEHS), and a number of other establishments from Research Triangle Park and beyond. Presentations ranged from the technology behind array production through the sharing of actual experimental data and projections on the future importance and applications of arrays. The information contained in the workshop presentations should provide aid and insight into arrays in general and their application to toxicology in particular.

Array Elements

In the context of molecular biology, the word "array" is normally used to refer to a series of DNA or protein elements firmly attached in a regular pattern to some kind of supportive medium. DNA array is often used interchangeably with gene array or microarray. Although not formally defined, microarray is generally used to describe the higher density arrays typically printed on glass chips. The DNA elements that make up DNA arrays can be oligonucleotides, partial gene sequences, or full-length cDNAs. Companies offering pre-made arrays that contain less than full-length clones normally use regions of the genes which are specific to that gene to prevent false positives arising through cross-hybridization. Sequence verification of cDNA clone identity is necessary because of errors in identifying specific clones from cDNA libraries and databases. Premade DNA arrays printed on membranes are currently or imminently available for human, mouse, and rat. In most cases they contain DNA sequences representing several thousand different sequence clusters or genes as delineated through the National Center for Biotechnology Information UniGene Project (*2*). Many of these different UniGene clusters (putative genes) are represented only by expressed sequence tags (ESTs).

Array Printing

Arrays are typically printed on one of two types of support matrix. Nylon membranes are used by most off-the-shelf array providers such as Clontech Laboratories, Inc. (Palo Alto, CA), Genome Systems, Inc. (St. Louis, MO), and Research Genetics, Inc. (Huntsville, AL). Microarrays such as those produced by Affymetrix, Inc. (Santa Clara, CA), Incyte Pharmaceuticals, Inc. (Palo Alto, CA), and many do-it-yourself (DIY) arraying groups use glass wafers or slides. Although standard microscope slides may be used, they must be preprepared to facilitate sticking of the DNA to the glass. Several different

Address correspondence to J. Rockett, Reproductive Toxicology Division (MD-72), National Health and Environmental Effects Research Laboratory, U.S. EPA, Research Triangle Park, NC 27711 USA. Telephone: (919) 541-2678. Fax: (919) 541-4017. E-mail: rockett.john@epa.gov

The authors thank R. Kavlock for envisioning the application of array technology to toxicology at the U.S. Environmental Protection Agency. We also thank T. Wall and B. Deitz for administrative assistance.

This document has been reviewed in accordance with EPA policy and approved for publication. Mention of companies, trade names, or products does not signify endorsement of such by the EPA.

Received 23 March 1999; accepted 22 April 1999.

coatings have been successfully used, including silane and lysine. The coating of slides can easily be carried out in the laboratory, but many prefer the convenience of precoated slides available from suppliers.

Once the support matrix has been prepared, the DNA elements can be applied by several methods. Affymetrix, Inc., has developed a unique photolithographic technology for attaching oligonucleotides to glass wafers. More commonly, DNA is applied by either noncontact or contact printing. Noncontact printers can use thermal, solenoid, or piezoelectric technology to spray aliquots of solution onto the support matrix and may be used to produce slide or membrane-based arrays. Cartesian Technologies, Inc. (Irvine, CA) has developed nQUAD technology for use in its PixSys printers. The system couples a syringe pump with the microsolenoid valve, a combination that provides rapid quantitative dispensing of nanoliter volumes (down to 4.2 nL) over a variable volume range. A different approach to noncontact printing uses a solid pin and ring combination (Genetic MicroSystems, Inc., Woburn, MA). This system (Figure 1) allows a broader range of sample, including cell suspensions and particulates, because the printing head cannot be blocked up in the same way as a spray nozzle. Fluid transfer is controlled in this system primarily by the pin dimensions and the force of deposition, although the nature of the support matrix and the sample will also affect transfer to some degree.

In contact printing, the pin head is dipped in the sample and then touched to the support matrix to deposit a small aliquot. Split pins were one of the first contact-printing devices to be reported and are the suggested format for DIY arrayers, as described by Brown (*3*). Split pins are small metal pins with a precise groove cut vertically in the middle of the pin tip. In this system, 1–48 split pins are positioned in the pin-head. The split pins work by simple capillary action, not unlike a fountain pen—when the pin heads are dipped in the sample, liquid is drawn into the pin groove. A small (fixed) volume is then deposited each time the split pins are gently touched to the support matrix. Sample (100–500 pL depending on a variety of parameters) can be deposited on multiple slides before refilling is required, and array densities of > 2,500 spots/cm^2 may be produced. The deposit volume depends on the split size, sample fluidity, and the speed of printing. Split pins are relatively simple to produce and can be made in-house if a suitable machine shop is available. Alternatively, they can be obtained directly from companies such as TeleChem International, Inc. (Sunnyvale, CA).

Irrespective of their source, printers should be run through a preprint sequence prior to producing the actual experimental arrays; the first 100 or so spots of a new run tend to be somewhat variable. Factors effecting spot reproducibility include slide treatment homogeneity, sample differences, and instrument errors. Other factors that come into play include clean ejection of the drop and clogging (nQUAD printing) and mechanical variations and long-term alteration in print-head surface of solid and split pins. However, with careful preparation it is possible to get a coefficient of variance for spot reproducibility below 10%.

One potential printing problem is sample carryover. Repeated washing, blotting, and drying (vacuum) of print pins between samples is normally effective at reducing sample carryover to negligible amounts. Printing should also be carried out in a controlled environment. Humidified chambers are available in which to place printers. These help prevent dust contamination and produce a uniform drying rate, which is important in determining spot size, quality, and reproducibility.

In summary, although several printing technologies are available, none are particularly outstanding and the bottom line is that they are still in a relatively early stage of evolution.

Array Hybridization

The hybridization protocol is, practically speaking, relatively straightforward and those with previous experience in blotting should have little difficulty. Array hybridizations are, in essence, reverse Southern/Northern blots—instead of applying a labeled probe to the target population of DNA/RNA, the labeled population is applied to the probe(s). With membrane-based arrays, the control and treated mRNA populations are normally converted to cDNA and labeled with isotope (e.g., ^{33}P) in the process. These labeled populations are then hybridized independently to parallel or serial arrays and the hybridization signal is detected with a phosporimager. A less commonly used alternative to radioactive probes is enzymatic detection. The probe may be biotinylated, haptenylated, or have alkaline phosphatase/horseradish peroxidase attached. Hybridization is detected by enzymatic reaction yielding a color reaction (*4*). Differences in hybridization signals can be detected by eye or, more accurately, with the help of digital imaging and commercially available software. The labeling of the test populations for slide-based microarrays uses a slightly different approach. The probe typically consists of two samples of polyA$^+$ RNA (usually from a treated and a control population) that are converted to cDNA; in the process each is labeled with a different fluor. The independently labeled probes are then mixed together and hybridized to a single microarray slide and the resulting combined fluorescent signal is scanned. After normalization, it is possible to determine the ratio of fluorescent signals from a single hybridization of a slide-based microarray.

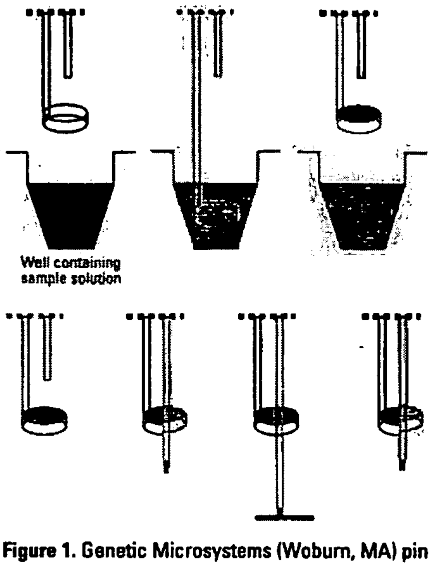


Figure 1. Genetic Microsystems (Woburn, MA) pin ring system for printing arrays. The pin ring combination consists of a circular open ring oriented parallel to the sample solution, with a vertical pin centered over the ring. When the ring is dipped into a solution and lifted, it withdraws an aliquot of sample held by surface tension. To spot the sample, the pin is driven down through the ring and a portion of the solution is transferred to the bottom of the pin. The pin continues to move downward until the pendant drop of solution makes contact with the underlying surface. The pin is then lifted, and gravity and surface tension cause deposition of the spot onto the array. Figure from Flowers et al. (*14*), with permission from Genetic Microsystems.

cDNA derived from control and treated populations of RNA is most commonly hybridized to arrays, although subtractive hybridization or differential display reactions may also be used. Fluorophore- or radiolabeled nucleotides are directly incorporated into the cDNA in the process of converting RNA to cDNA. Alternatively, 5′ end-labeled primers may be used for cDNA synthesis. These are labeled with a fluorophore for direct visualization of the hybridized array. Alternatively, biotin or a hapten may be attached to the primer, in which case fluor-labeled streptavidin or antibody must be applied before a signal can be generated. The most commonly used fluorophores at present are cyanine (Cy)3 and Cy5 (Amersham Pharmacia Biotech AB, Uppsala, Sweden). However, the relative expense of these fluorescent conjugates has driven a search for cheaper alternatives. Fluorescein, rhodamine, and Texas red have all been used, and companies such as Molecular Probes, Inc. (Eugene, OR) are developing a series of labeled nucleotides with a wide range of excitation and emission spectra which may prove to function as well as the Cy dyes.

Analysis of DNA Microarrays

Membrane-based arrays are normally analyzed on film or with a phosphorimager, whereas chip-based arrays require more specialized scanning devices. These can be divided into three main groups: the charge-coupled device camera systems, the nonconfocal laser scanners, and the confocal laser scanners. The advantages and disadvantages of each system are listed in Table 1.

Table 1. Advantages and disadvantages of different microarray scanning systems.

		Nonconfocal laser scanner	
Advantages	Few moving parts	Relatively simple optics	Small depth of focus reduces artifacts
	Fast scanning of bright samples		May have high light collection efficiency
Disadvantages	Less appropriate for dim samples	Low light collection efficiency	Small depth of focus requires scanning precision
	Optical scatter can limit performance	Background artifacts not rejected	
		Resolution typically low	

CCD, charge-coupled device.
From Kawasaki (*13*).

Because a typical spot on a microarray can contain > 10^8 molecules, it is clear that a large variation in signal strength may occur. Current scanners cannot work across this many orders of magnitude (4 or 5 is more typical). However, the scanning parameters can normally be adjusted to collect more or less signal, such that two or three scans of the same array should permit the detection of rare and abundant genes.

When a microarray is scanned, the fluorescent images are captured by software normally included with the scanner. Several commercial suppliers provide additional software for quantifying array images, but the software tools are constantly evolving to meet the developing needs of researchers, and it is prudent to define one's own needs and clarify the exact capabilities of the software before its purchase. Issues that should be considered include the following:

- Can the software locate offset spots?
- Can it quantitate across irregular hybridization signals?
- Can the arrayed genes be programmed in for easy identification and location?
- Can the software connect via the Internet to databases containing further information on the gene(s) of interest?

One of the key issues raised at the workshop was the sensitivity of microarray technology. Experiments by General Scanning, Inc. (Watertown, MA), have shown that by using the Cy dyes and their scanner, signal can be detected down to levels of < 1 fluor molecule per square micrometer, which translates to detecting a rare message at approximately one copy per cell or less.

Array Applications

Although arrays are an emerging technology certain to undergo improvement and alteration, they have already been applied usefully to a number of model systems. Arrays are at their most powerful when they contain the entire genome of the species they are being used to study. For this reason, they have strong support among researchers utilizing yeast and *Caenorhabditis elegans* (*5*). The genomes of both of these species have been sequenced and, in the case of yeast, deposited onto arrays for examination of gene expression (*6,7*). With both of these species, it is relatively easy to perturb individual gene expression. Indeed, *C. elegans* knockouts can be made simply by soaking the worms in an antisense solution of the gene to be knocked out.

By a process of systematic gene disruption, it is now possible to examine the cause and effect relationships between different genes in these simple organisms. This kind of approach should help elucidate biochemical pathways and genetic control processes, deconvolute polygenic interactions, and define the architecture of the cellular network. A simple case study of how this can be achieved was presented by Butow [University of Texas Southwestern Medical Center, Dallas, TX (Figure 2)]. Although it is the phenotypic result of a single gene knockout that is being examined, the effect of such perturbation will almost always be polygenic. Polygenic interactions will become increasingly important as researchers begin to move away from single gene systems when examining the nature of toxicologic responses to external stimuli. This is especially important in toxicology because the phenotype produced by a given environmental insult is never the result of the action of a single gene; rather, it is a complex interaction of one or multiple cellular pathways. Phenomena such as quantitative trait (the continuous variation of phenotype), epistasis (the effect of alleles of one or more genes on the expression of other genes), and penetrance (proportion of individuals of a given genotype that display a particular phenotype) will become increasingly evident and important as toxicologists push toward the ultimate goal of matching the responses of individuals to different environmental stimuli.

Analysis of the transcriptome (the expression level of all the genes in a given cell population) was a use of arrays addressed by several speakers. Unfortunately, current gene nomenclature is often confusing in that single genes are allocated multiple names (usually as a result of independent discovery by different laboratories), and there was a call for standardization of gene nomenclature. Nevertheless, once a transcriptome has been assembled it can then be transferred onto arrays and used to screen any chosen system. The EPA MicroArray Consortium (EPAMAC) is assembling testes transcriptomes for human, rat, and mouse. In a slightly different approach, Nuwaysir et al. (*8*) describes how the NIEHS assembled what is effectively a "toxicological transcriptome"—a library of human and mouse genes that have previously been proven or implicated in responses to toxicologic insults. Clontech Laboratories, Inc. (Palo Alto, CA), has begun a similar process by developing stress/toxicology filter arrays of rat, mouse, and human genes. Thus, rather than being tissue or cell specific, these stress/toxicology arrays can be used across a variety of model systems to look for alterations in the expression of toxicologically important genes and define the new field of toxicogenomics. The potential to identify toxicant families based on tissue- or cell-specific gene expression could revolutionize drug testing. These molecular signatures or fingerprints could not only point to the possible toxicity/carcinogenicity of newly discovered compounds (Figure 3), but also aid in elucidating their mechanism of action through identification of gene expression networks. By extension, such signatures could provide easily identifiable biomarkers to assess the degree, time, and nature of exposure.

DNA arrays are primarily a tool for examining differential gene expression in a given model. In this context they are referred to as closed systems because they lack the ability of other differential expression technologies, e.g., differential display and subtractive hybridization, to detect previously unknown genes not present on the array. This would appear to limit the power of DNA arrays to the imaginations and preconceptions of the researcher in selecting genes previously characterized and thought to be involved in the model system. However, the various genome sequencing projects have created a new category of sequence—the EST—that has partially mollified this deficiency. ESTs are cDNAs expressed in a given tissue that, although they may share some degree of sequence similarity to previously characterized genes, have not been assigned specific genetic identity. By incorporating EST clones into an array, it is possible to monitor the expression of these unknown genes. This can enable the identification of previously uncharacterized genes that may have biologic significance in the model system. Filter arrays from Research Genetics and slide arrays from Incyte Pharmaceuticals both incorporate large numbers of ESTs from a variety of species.

A further use of microarrays is the identification of single nucleotide polymorphisms (SNPs). These genomic variations are abundant—they occur approximately every 1 kb or so—and are the basis of restriction fragment length polymorphism analysis used in forensic analysis. Affymetrix, Inc., designed chips that contain multiple repeats of the same gene sequence. Each position is present with all four possible bases. After the hybridization of the sample, the degree of hybridization to the different sequences can be measured and the exact sequence of the target gene deduced. SNPs are thought to be of vital importance in drug metabolism and toxicology. For example, single base differences in the regulatory region or active site of some genes can account for huge differences in the activity of that gene. Such SNPs are thought to explain why some people are able to metabolize certain xenobiotics better than others. Thus, arrays provide a further tool for the toxicologist investigating the nature of susceptible subpopulations and toxicologic response.

There are still many wrinkles to be ironed out before arrays become a standard tool for toxicologists. The main issues raised at the workshop by those with hands-on experience were the following:

- Expense: the cost of purchasing/contracting this technology is still too great for many individual laboratories.
- Clones: the logistics of identifying, obtaining, and maintaining a set of nonredundant, noncontaminated, sequence-verified, species/cell/tissue/field-specific clones.
- Use of inbred strains: where whole-organism models are being used, the use of inbred strains is important to reduce the potentially confusing effects of the individual variation typically seen in outbred populations.
- Probe: the need for relatively large amounts of RNA, which limits the type of sample (e.g., biopsy) that can be used. Also, different RNA extraction methods can give different results.
- Specificity: the ability to discriminate accurately between closely related genes (e.g., the cytochrome p450 family) and splice variants.
- Quantitation: the quantitation of gene expression using gene arrays is still open to debate. One reason for this is the different incorporation of the labeling dyes. However, the main difficulty lies in knowing what to normalize against. One option is to include a large number of so-called housekeeping genes in the array. However, the expression of these genes often change depending on the tissue and the toxicant, so it is necessary to characterize the expression of these genes in the model system before utilizing them. This is clearly not a viable option when screening multiple new compounds. A second option is to include on the array genes from a nonrelated species (e.g., a plant gene on an animal array) and to spike the probe with synthetic RNA(s) complementary to the gene(s).
- Reproducibility: this is sometimes questionable, and a figure of approximately two or three repeats was used as the minimum number required to confirm initial findings. Again, however, most people advocated the use of Northern blots or reverse transcriptase PCR to confirm findings.
- Sensitivity: concerns were voiced about the number of target molecules that must be present in a sample for them to be detected on the array.
- Efficiency: reproducible identification of 1.5- to 2-fold differences in expression was reported, although the number of genes that undergo this level of change and remain undetected is open to debate. It is important that this level of detection be ultimately achieved because it is commonly perceived that some important transcription factors and their regulators respond at such low levels. In most cases, 3- to 5-fold was the minimum change that most were happy to accept.
- Bioinformatics: perhaps the greatest concern was how to accurately interpret the data with the greatest accuracy and efficiency. The biggest headache is trying to identify networks of gene expression that are common to different treatments or doses. The amount of data from a single experiment is huge. It may be that, in the future, several groups individually equipped with specialized software algorithms for studying their favorite genes or gene systems will be able to share the same hybridized chips. Thus, arrays could usher in a new perspective on collaboration and the sharing of data.

EPAMAC

Perhaps the main reason most scientists are unable to use array technology is the high cost involved, whether buying off-the-shelf membranes, using contract printing services, or

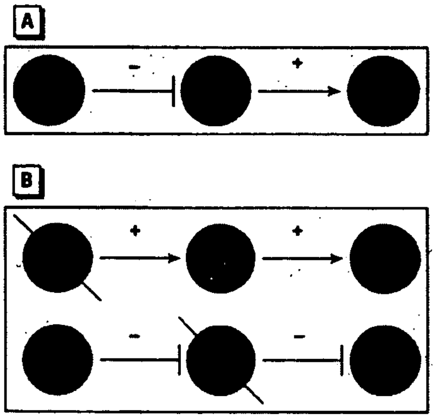
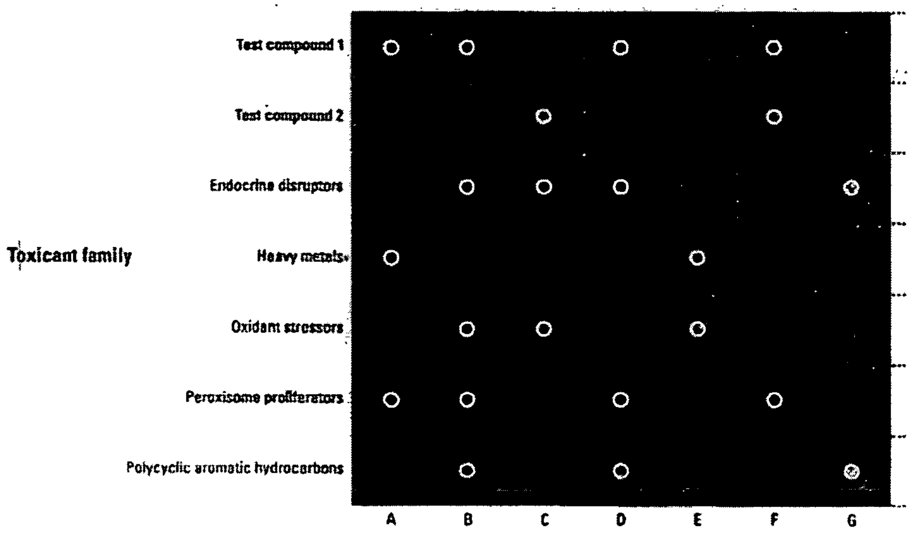


Figure 2. Potential effects of gene knockout within positively and negatively regulated gene expression networks. i_1 is limiting in wild type for expression of i_2. (*A*) A simple, two-component, linear regulatory network operating on gene i_2, where i_1 is a positive effector of i_2 and j_n is either a positive or negative effector of i_1. This network could be deduced by examining the consequence of (*B*) deleting j_n on the expression of i_1 and i_2, where the expression of i_2 would be decreased or increased depending on whether j_n was a positive or negative regulator. These and other connected components of even greater complexity could be revealed by genome-wide expression analysis. From Butow (*15*).



Toxicant family	A	B	C	D	E	F	G
Test compound 1	○	○		○		○	
Test compound 2			○			○	
Endocrine disruptors		○	○	○			○
Heavy metals	○				○		
Oxidant stressors		○	○		○		
Peroxisome proliferators	○	○		○		○	
Polycyclic aromatic hydrocarbons		○		○			○

Figure 3. Gene expression profiles—also called fingerprints or signatures—of known toxicants or toxicant families may, in the future, be used to identify the potential toxicity of new drugs, etc. In this example, the genetic signature of test compound 1 is identical to that of known peroxisome proliferators, whereas that of test compound 2 does not match any known toxicant family. Based on these results, test compound 2 would be retained for further testing and test compound 1 would be eliminated.

producing chips in-house. In view of this, researchers at the RTD/NHEERL initiated the EPAMAC. This consortium brings together scientists from the EPA and a number of extramural labs with the aim of developing microarray capability through the sharing of resources and data. EPAMAC researchers are primarily interested in the developmental and toxicologic changes seen in testicular and breast tissue, and a portion of the workshop was set aside for EPAMAC members to share their ideas on how the experimental application of microarrays could facilitate their research. One of the central areas of interest to EPAMAC members is the effect of xenobiotics on male fertility and reproductive health. Of greatest concern is the effect of exposure during critical periods of development and germ cell differentiation (*9*), and how this may compromise sperm counts and quality following sexual maturation (*10*). As well as spermatogenic tissue, there is also interest in how residual mRNA found in mature sperm (*11*) could be used as an indicator of previous xenobiotic effects (it is easier to obtain a semen sample than a testicular biopsy). Arrays will be used to examine and compare the effect of exposure to heat and chemicals in testicular and epididymal gene expression profiles, with the aim of establishing relationships/associations between changes in developmental landmarks and the effects on sperm count and quality. Cluster, pattern, and other analysis of such data should help identify hidden relationships between genes that may reveal potential mechanisms of action and uncover roles for genes with unknown functions.

Summary

The full impact of DNA arrays may not be seen for several years, but the interest shown at this regional workshop indicates the high level of interest that they foster. Apart from educating and advertising the various technologies in this field, this workshop brought together a number of researchers from the Research Triangle Park area who are already using DNA arrays. The interest in sharing ideas and experiences led to the initiation of a Triangle array user's group.

Array technology is still in its infancy. This means that the hardware is still improving and there is no current consensus for standard procedures, quantitation, and interpretation. Consistency in spotting and scanning arrays is not yet optimized, and this is one of the most critical requirements of any experiment. In addition, one of the dark regions of array technology—strife in the courts over who owns what portions of it—has further muddled the future and is a potential barrier toward the development of consensus procedures.

Perhaps the greatest hurdle for the application of arrays is the actual interpretation of data. No specialists in bioinformatics attended the workshop, largely because they are rare and because as yet no one seems clear on the best method of approaching data analysis and interpretation. Cross-referencing results from multiple experiments (time, dose, repeats, different animals, different species) to identify commonly expressed genes is a great challenge. In most cases, we are still a long way from understanding how the expression of gene *X* is related to the expression of gene *Y*, and ordering gene expression to delineate causal relationships.

To the ordinary scientist in the typical laboratory, however, the most immediate problem is a lack of affordable instrumentation. One can purchase premade membranes at relatively affordable prices. Although these may be useful in identifying individual genes to pursue in more detail using other methods, the numbers that would be required for even a small routine toxicology experiment prohibit this as a truly viable approach. For the toxicologist, there is a need to carry out multiple experiments—dose responses, time curves, multiple animals, and repeats. Glass-based DNA arrays are most attractive in this context because they can be prepared in large batches from the same DNA source and accommodate control and treated samples on the same chip. Another problem with current off-the-shelf arrays is that they often do not contain one or more of the particular genes a group is interested in. One alternative is to obtain and/or produce a set of custom clones and have contract printing of membranes or slides carried out by a company such as Genomic Solutions, Inc. (Ann Arbor, MI). This approach is less expensive than laying out capital for one's own entire system, although at some point it might make economic sense to print one's own arrays.

Finally, DNA arrays are currently a team effort. They are a technology that uses a wide range of skills including engineering, statistics, molecular biology, chemistry, and bioinformatics. Because most individuals are skilled in only one or perhaps two of these areas, it appears that success with arrays may be best expected by teams of collaborators consisting of individuals having each of these skills.

Those considering array applications may be amused or goaded on by the following quote from *Fortune* magazine (*12*):

> Microprocessors have reshaped our economy, spawned vast fortunes and changed the way we live. Gene chips could be even bigger.

Although this comment may have been designed to excite the imagination rather than accurately reflect the truth, it is fair to say that the age of functional genomics is upon us. DNA arrays look set to be an important tool in this new age of biotechnology and will likely contribute answers to some of toxicology's most fundamental questions.

Speakers

Cindy Afshari
NIEHS
Linda Birnbaum
U.S. EPA
Ron Butow
University of Texas
Southwestern Medical
Center
Alex Chenchik
Clontech Laboratories, Inc.
David Dix
U.S. EPA
Abdel Elkahloun
Research Genetics, Inc.
Sue Fenton
U.S. EPA
Norman Hecht
University of Pennsylvania
Pat Hurban
Paradigm Genetics, Inc.
Bob Kavlock
U.S. EPA
Ernie Kawasaki
General Scanning, Inc.
Steve Krawetz
Wayne State University
Nick Mace
Genetic Microsystems, Inc.
Scott Mordecai
Affymetrix, Inc.
Kevin Morgan
Glaxo Wellcome, Inc.
Elaine Poplin
Research Genetics, Inc.
Don Rose
Cartesian Technologies, Inc.
Jim Samet
U.S. EPA
Sam Ward
University of Arizona
Jeff Welch
U.S. EPA
Reen Wu
University of California
at Davis
Tim Zacharewski
Michigan State University

References and Notes

1. The chipping forecast. Nat Genet 21(Suppl 1):3–60 (1999).
2. National Center for Biotechnology Information. The Unigene System. Available: www.ncbi.nlm.nih.gov/Schuler/UniGene [cited 22 March 1999].
3. Brown PO. The Brown Lab. Available: http://cmgm.Stanford.edu/pbrown [cited 22 March 1999].
4. Chen JJ, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF, Cheng F, et al. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics 51:313–324 (1998).
5. Ward S. DNA Microarray Technology to Identify Genes Controlling Spermatogenesis. Available: www.mcb.arizona.edu/wardlab/microarray.html [cited 22 March 1999].
6. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 4:1293–1301 (1998).
7. Brown PO. The Full Yeast Genome on a Chip. Available: http://cmgm.stanford.edu/pbrown/yeastchip.html [cited 22 March 1999].
8. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog 24(3):153–159 (1999).
9. Hecht NB. Molecular mechanisms of male germ cell differentiation. Bioessays 20:555–561 (1998).
10. Zacharewski TR. Timothy R. Zacharewski. Available: www.bch.msu.edu/faculty/zachar.htm [cited 22 March 1999].
11. Kramer JA, Krawetz SA. RNA in spermatozoa: implications for the alternative haploid genome. Mol Hum Reprod 3:473–478 (1997).
12. Stipp D. Gene chip breakthrough. Fortune, March 31:56–73 (1997).
13. Kawasaki E (General Scanning Instruments, Inc., Watertown, MA). Unpublished data.
14. Flowers P, Overbeck J, Mace ML Jr, Pagliughi FM, Eggers WJE, Yonkers H, Honkanen P, Montagu J, Rose SD. Development and Performance of a Novel Microarraying System Based on Surface Tension Forces. Available: http://www.geneticmicro.com/resources/html/coldspring.html [cited 22 March 1999].
15. Butow R (University of Texas Medical Center, Dallas, TX). Unpublished data.

Reference 5 of 9
with Response dated 4/23/04
In USSN: 10/049,742

Subject: RE: [Fwd: Toxicology Chip]
Date: Mon. 3 Jul 2000 08:09:45 -0400
From: "Afshari.Cynthia" <afshari@niehs.nih.gov>
To: "'Diana Hamlet-Cox'" <dianahc@incyte.com>

```
You can see the list of clones that we have on our 12K chip at
http://manuel.niehs.nih.gov/maps/guest/clonesrch.cfm
We selected a subset of genes (2000K) that we believed critical to tox
response and basic cellular processes and added a set of clones and ESTs to
this.  We have included a set of control genes (80+) that were selected by
the NHGRI because they did not change across a large set of array
experiments.  However, we have found that some of these genes change
signficantly after tox treatments and are in the process of looking at the
variation of each of these 80+ genes across our experiments.
Our chips are constantly changing and being updated and we hope that our
data will lead us to what the toxchip should really be.
I hope this answers your question.
Cindy Afshari

> ----------
> From:          Diana Hamlet-Cox
> Sent:          Monday, June 26, 2000 8:52 PM
> To:   afshari@niehs.nih.gov
> Subject:       [Fwd: Toxicology Chip]
>
> Dear Dr. Afshari,
>
> Since I have not yet had a response from Bill Grigg, perhaps he was not
> the right person to contact.
>
> Can you help me in this matter?  I don't need to know the sequences,
> necessarily, but I would like very much to know what types of sequences
> are being used, e.g., GPCRs (more specific?), ion channels, etc.
>
> Diana Hamlet-Cox
>
> -------- Original Message --------
> Subject: Toxicology Chip
> Date: Mon, 19 Jun 2000 18:31:48 -0700
> From: Diana Hamlet-Cox <dianahc@incyte.com>
> Organization: Incyte Pharmaceuticals
> To: grigg@niehs.nih.gov
>
> Dear Colleague:
>
> I am doing literature research on the use of expressed genes as
> pharmacotoxicology markers, and found the Press Release dated February
> 29, 2000 regarding the work of the NIEHS in this area.  I would like to
> know if there is a resource I can access (or you could provide?) that
> would give me a list of the 12,000 genes that are on your Human ToxChip
> Microarray.  In particular, I am interested in the criteria used to
> select sequences for the ToxChip, including any control sequences
> included in the microarray.
>
> Thank you for your assistance in this request.
>
> Diana Hamlet-Cox, Ph.D.
> Incyte Genomics, Inc.
>
> --
>
> =======================
```

> This email message is for the sole use of the intended recipient(s) and
> may contain confidential and privileged information subject to
> attorney-client privilege. Any unauthorized review, use, disclosure or
> distribution is prohibited. If you are not the intended recipient,
> please contact the sender by reply email and destroy all copies of the
> original message.
> ==========================
>
>

Reference 6 of 9
with Response dated 4/23/04
In USSN: 10/049,742

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT **Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.**

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

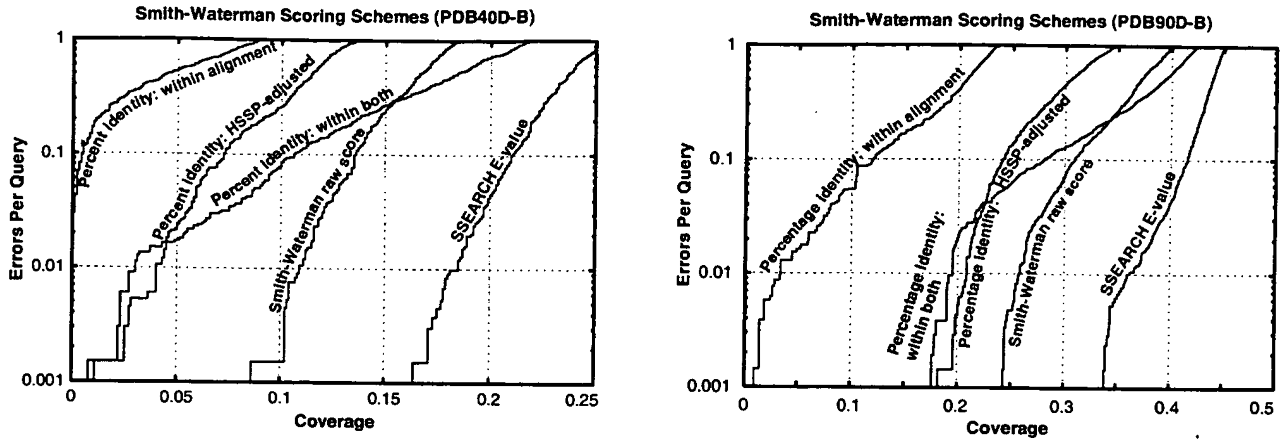
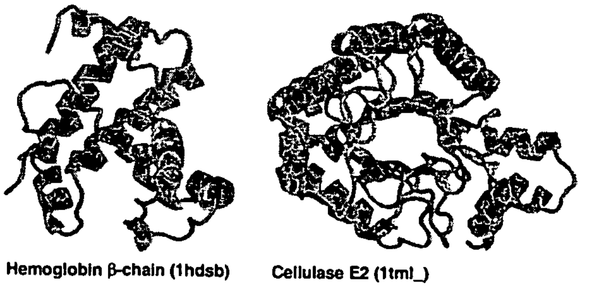


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith–Waterman. (*A*) Analysis of PDB40D-B database. (*B*) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the *x* axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The *y* axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The *y* axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith–Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Reciever Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely



```
1hdsb GKVDVDVVGAQALGR--LLVVYPWTQRFFQHFGNLSSAGAVMNNPKVKAHGKRVLDAFTQGLKH
      :.::. . .::: :.  .::::    :     :: ::.::    :. .:. .  .: :. :::.
1tml_ GQVDALMSAAQAAGKIPILVVYNAPGR---DCGNHSSGGA----PSHSAY-RSWIDEFAAGLKN
```

FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin β-chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

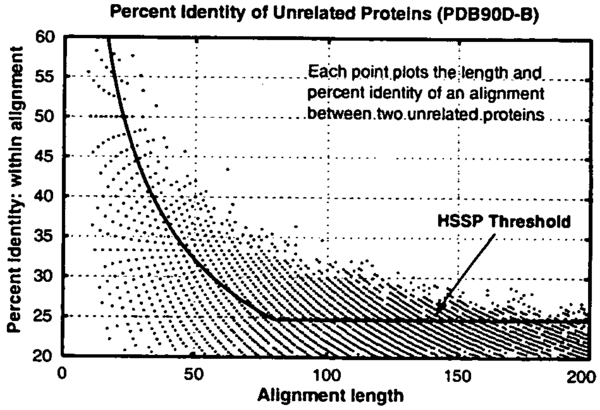


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5*A* and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5*B*). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

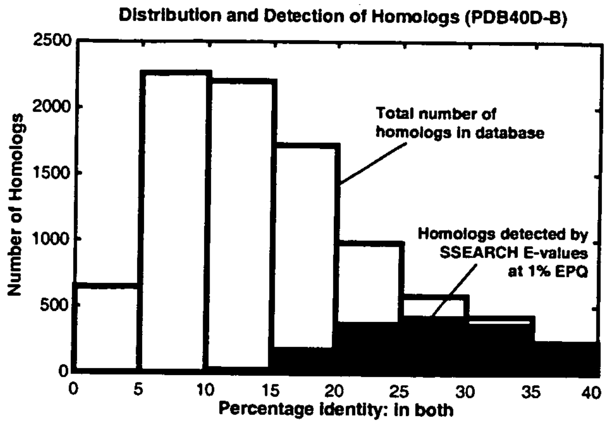


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pariwise sequence comparison to detect them.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (*i*) using a large current database in which the protein sequences have been complexity masked and (*ii*) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

Proteomics: a major new technology for the drug discovery process

Martin J. Page, Bob Amess, Christian Rohlff, Colin Stubberfield and Raj Parekh

Proteomics is a new enabling technology that is being integrated into the drug discovery process. This will facilitate the systematic analysis of proteins across any biological system or disease, forwarding new targets and information on mode of action, toxicology and surrogate markers. Proteomics is highly complementary to genomic approaches in the drug discovery process and, for the first time, offers scientists the ability to integrate information from the genome, expressed mRNAs, their respective proteins and subcellular localization. It is expected that this will lead to important new insights into disease mechanisms and improved drug discovery strategies to produce novel therapeutics.

Among the major pharmaceutical and biotechnology companies, it is clearly recognized that the business of modern drug discovery is a highly competitive process. All of the many steps involved are inherently complex, and each can involve a high risk of attrition. The players in this business strive continuously to optimize and streamline the process; each seeking to gain an advantage at every step by attempting to make informed decisions at the earliest stage possible. The desired outcome is to accelerate as many key activities in the drug discovery process as possible. This should produce a new generation of robust drugs that offer a high probability of success and reach the clinic and market ahead of the competition.

There has been noticeable emphasis over recent years for companies to aggressively review and refine their strategies to discover new drugs. Central to this has been the introduction and implementation of cutting-edge technologies. Most, if not all, companies have now integrated key technology platforms that incorporate genomics, mRNA expression analysis, relational databases, high-throughput robotics, combinatorial chemistry and powerful bioinformatics. Although it is still early days to quantify the real impact of these platforms in clinical and commercial terms, expectations are high, and it is widely accepted that significant benefits will be forthcoming. This is largely based on data obtained during preclinical studies where the genomic[1,2] and microarray[3,4] technologies have already proved their value.

However, there are several noteworthy outcomes that result from this. Many comments are voiced that scientists armed with these technologies are now commonly faced with data overload. Thus, in some instances, rather than facilitating the decision process, the accumulation of more complex data points, many with unknown consequences, can seem to hinder the process. Also, most drug companies have simultaneously incorporated very similar components of the new technology platforms, the consequence being that it is becoming difficult yet again to determine where a clear competitive advantage will arise. Finally, in recent years, largely as a result of the accessibility of the technologies, there has been an overwhelming emphasis placed on genomic and mRNA data rather than on protein

Martin J. Page*, **Bob Amess**, **Christian Rohlff**, **Colin Stubberfield** and **Raj Parekh**, Oxford GlycoSciences, 10 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire, UK OX14 3YS. *tel: +44 1235 543277, fax: +44 1235 543283, e-mail: martin.page@ogs.co.uk

1359-6446/99/$ – see front matter © Elsevier Science. All rights reserved. PII: S1359-6446(98)01291-4

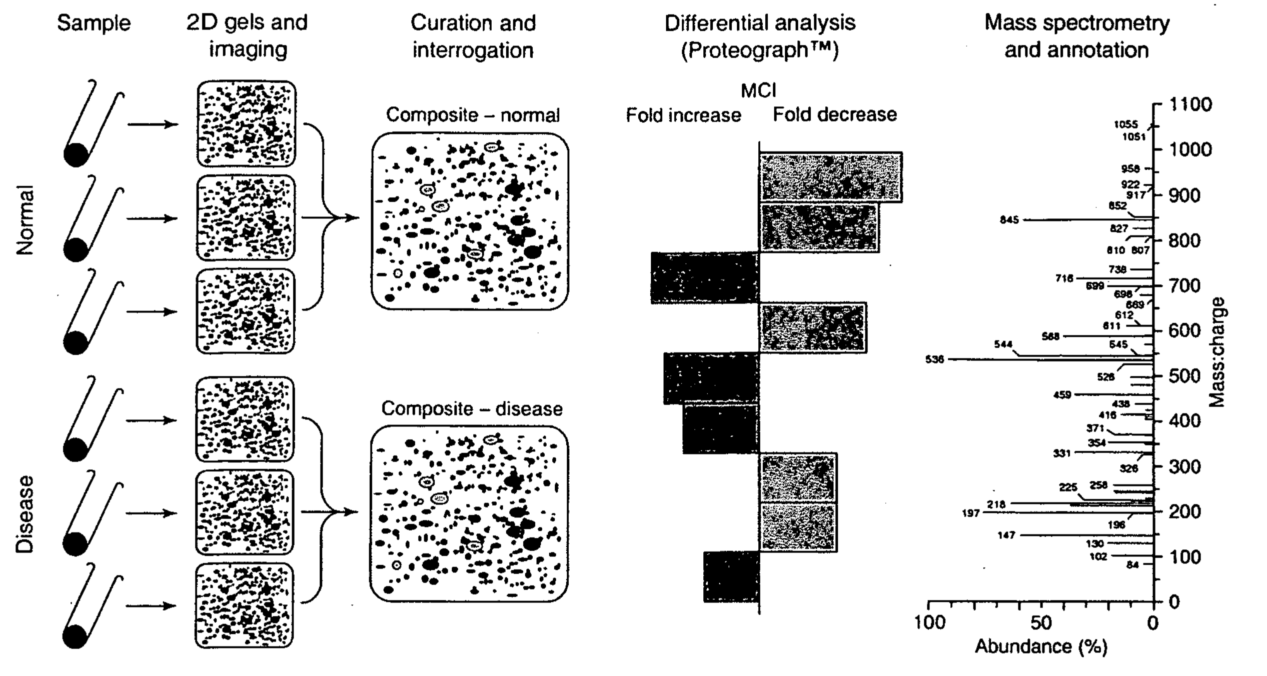


Figure 1. Steps involved in analysing a biological sample by proteomics. MCI, molecular cluster index.

analysis. It is important to remember that proteins dictate biological phenotype – whether it is normal or diseased – and are the direct targets for most drugs.

Proteomics: new technology for the analysis of proteins

It is now timely to recognize that complementary technology in the form of high-throughput analysis of the total protein repertoire of chosen biological samples, namely proteomics, is poised to add a new and important dimension to drug discovery. In a similar fashion to genomics, which aims to profile every gene expressed in a cell, proteomics seeks to profile every protein that is expressed[5–7]. However, there is added information, since proteomics can also be used to identify the post-translational modifications of proteins[8], which can have profound effects on biological function, and their cellular localization. Importantly, proteomics is a technology that integrates the significant advances in two-dimensional (2D) electrophoretic separation of proteins, mass spectrometry and bioinformatics. With these advances it is now possible to consistently derive proteomes that are highly reproducible and suitable for interrogation using advanced bioinformatic tools.

There are many variations whereby different laboratories operate proteomics. For the purpose of this review, the process used at Oxford GlycoSciences (OGS), which uses an industrial-scale operation that is integral to its drug discovery work, will be described. The individual steps of this process, where up to 1000 2D gels can be run and analysed per week, are summarized in Fig. 1. The incoming samples are bar coded and all information relevant to the sample is logged into a Laboratory Information Management System (LIMS) database. There can be a wide range in the type of samples processed, as applicable to individual steps in the drug discovery pipeline, and these will be mentioned later. The samples are separated according to their charge (pI) in the first dimension, using isoelectric focusing, followed by size (MW) using SDS–PAGE in the second dimension. Many modifications have been made to these steps to improve handling, throughput and reproducibility. The separated proteins are then stained with fluorescent dyes which are significantly more sensitive in detection than standard silver methods and have a broader dynamic range. The image of the displayed proteins obtained is referred to as the proteome, and is digitally scanned into databases using proprietary software called ROSETTA™. The images are subsequently curated, which begins with the removal of any artefacts, cropping and the placement of pI/MW landmarks. The images from replicate images are then aligned and matched to one another to generate a synthetic composite image. This is an important step, as the proteome is a dynamic situation, and it captures the biological variation that occurs, such that even orphan proteins are still incorporated into the analysis.

By means of illustration, Fig. 1 shows the process whereby proteomes are generated from normal and disease samples and how differentially expressed proteins are identified. The potential of this type of analysis is tremendous. For example, from a mammalian cell sample, in excess of 2000 proteins can typically be resolved within the proteome. The quality of this is shown in Fig. 2, which shows representative proteomes from three diverse biological sources: human serum, the pathogenic fungus *Candida albicans* and the human hepatoma cell line Huh7.

Use of proteomics to identify disease specific proteins

In most cases, the drug discovery process is initiated by the identification of a novel candidate target – almost always a protein – that is believed to be instrumental in the disease process. To date, there is a variety of means whereby drug targets have been forthcoming. These include molecular, cellular and genomic approaches, mostly centred upon DNA and mRNA analysis. The gene in question is isolated, and expression and characterization of its coded protein product – i.e. the drug target – is invariably a secondary event.

With the proteomic approach, the starting point is at the other end of the 'telescope'. Here there is direct and immediate comparison of the proteomes from paired normal and disease materials. Examples of these pairs are: (1) purified epithelial cell populations derived from human breast tumours, matched to purified normal populations of human breast epithelial cells, and (2) the invading pathogenic hyphal form of *C. albicans*, matched to the non-invading yeast form of *C. albicans*. When the proteome images from each pair are aligned, the Proteograph™ software is able to rapidly identify those proteins (each referenced as having a unique molecular cluster index, or MCI) that are either unique, or those that are differentially expressed. Thus, the Proteograph output from this analysis is both qualitative and quantitative.

Proteograph analysis for a particular study can also be undertaken on any number of samples. For example, one might compare anything from a few to several hundred preparations or samples, each from a normal and disease counterpart, and have these analysed in a single Proteograph study. In this way, it is possible to assign strong statistical confidence to the data and in some instances to identify specific subpopulations within the input biological sources. This feature will become increasingly significant in the near future, and there is a clear synergy here whereby proteomics can work closely with pharmacogenomic approaches to stratify patient populations and achieve effective targeted care for the patient. Whatever the source of the materials, the net output of Proteograph analysis is immediate identification of disease specific proteins. This is shown in Fig. 3, which shows the results of a proteograph obtained by comparing untreated human hepatoma cells with cells following exposure to a clinical

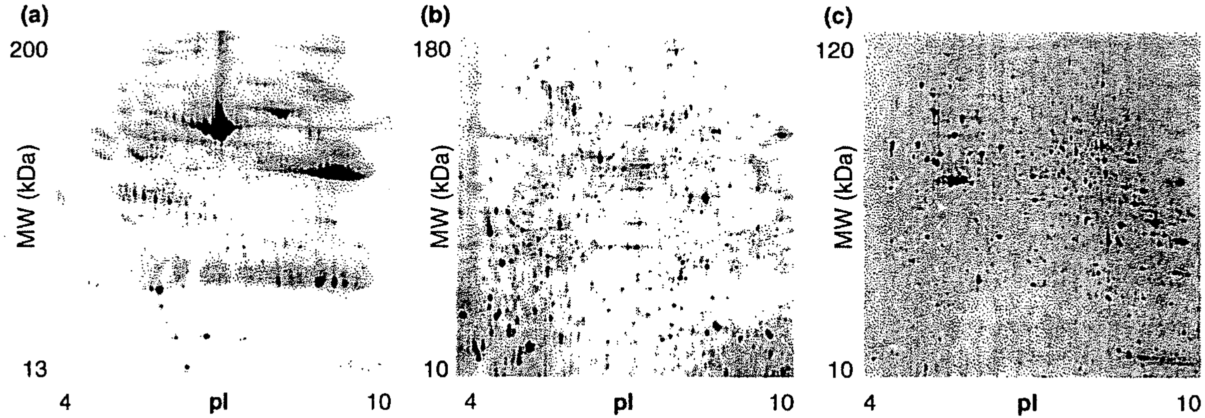
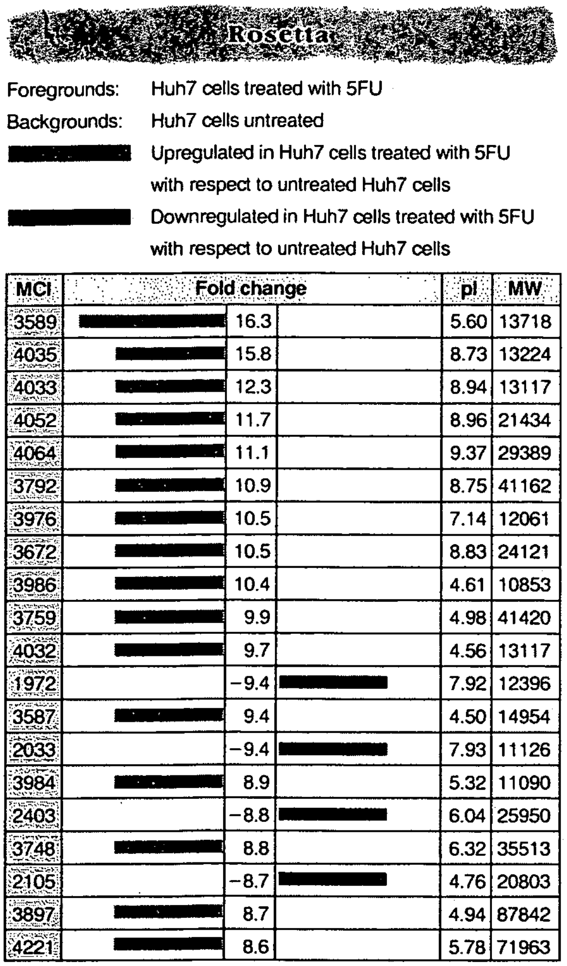


Figure 2. *Representative proteomes obtained from (a) human serum, (b) the pathogenic fungus* Candida albicans *and (c) the human hepatoma cell line Huh7.*



Foregrounds: Huh7 cells treated with 5FU

Backgrounds: Huh7 cells untreated

▬ Upregulated in Huh7 cells treated with 5FU with respect to untreated Huh7 cells

▬ Downregulated in Huh7 cells treated with 5FU with respect to untreated Huh7 cells

MCI	Fold change			pI	MW
3589	▬	16.3		5.60	13718
4035	▬	15.8		8.73	13224
4033	▬	12.3		8.94	13117
4052	▬	11.7		8.96	21434
4064	▬	11.1		9.37	29389
3792	▬	10.9		8.75	41162
3976	▬	10.5		7.14	12061
3672	▬	10.5		8.83	24121
3986	▬	10.4		4.61	10853
3759	▬	9.9		4.98	41420
4032	▬	9.7		4.56	13117
1972		−9.4	▬	7.92	12396
3587	▬	9.4		4.50	14954
2033		−9.4	▬	7.93	11126
3984	▬	8.9		5.32	11090
2403		−8.8	▬	6.04	25950
3748	▬	8.8		6.32	35513
2105		−8.7	▬	4.76	20803
3897	▬	8.7		4.94	87842
4221	▬	8.6		5.78	71963

Figure 3. *Table of differential protein expression profiles, referred to as a Rosetta Proteograph™, between Huh7 cells with and without the cytotoxic agent 5-FU. Bars are quantized and do not represent exact fold change values.*

cytotoxic agent. In this instance, only the top 20 differentially expressed MCIs are shown, but the readout would normally extend to a defined cut-off value, typically a two-fold or greater difference in expression levels, determined by the user.

In a typical analysis involving disease and normal mammalian material, in which each proteome would have ~2000 protein features each assigned an MCI, the proteograph might identify somewhere in the region of 50–300 MCIs that are unique or differentially expressed. To capitalize rapidly on these data, at OGS a high-throughput mass spectrometry facility coupled to advanced databases to annotate these MCIs as individual proteins is applied. As these are all disease specific proteins, each could represent a novel target and/or a novel disease marker. The process becomes even more powerful when a panel of features, rather than individual features, are assigned. The relevance of this is apparent when one considers that most diseases, if not all, are multifactorial in nature and arise from polygenic changes. Rather than analysing events in isolation, the ability to examine hundreds or thousands of events simultaneously, as shown by proteomics, can offer real advantages.

Identification and assignment of candidate targets

The rapid identification and assignment of candidate targets and markers represents a huge challenge, but this has been greatly facilitated by combining the recent advances made in proteomics and analytical mass spectrometry[9]. Using automated procedures it is now possible to annotate proteins present in femtomole quantities, which would depict the low abundance class of proteins. The process of annotation is similarly aided by the quality and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals). In this respect, the advances in proteomics have benefited considerably from the breakthroughs achieved with genomics.

From an application perspective, cancer studies provide a good opportunity whereby proteomics can be instrumental in identifying disease specific proteins, because it is often feasible to obtain normal and diseased tissue from the same patient. For example, proteomic studies have been reported on neuroblastomas[10], human breast proteins from normal and tumour sources[11–13], lung tumours[14], colon tumours[15] and bladder tumours[16]. There are also proteomic studies reported within the cardiovascular therapeutic area, in which disease or response proteins are identified[17,18].

Genomic microarray analysis can similarly identify unique species or clusters of mRNAs that are disease specific. However, in some instances, there is a clear lack of correlation between the levels of a specific mRNA and its corresponding protein (Ref. 19, Gypi, S.P. *et al.*, submitted). This has now been noted by many investigators and reaffirms that post-transcriptional events, including protein stability, protein modification (such as phosphorylation, glycosylation, acylation and methylation) and cell localization, can constitute major regulatory steps. Proteomic analysis captures all of these steps and can therefore provide unique and valuable information independent from, or complementary to, genomic data.

Proteomics for target validation and signal transduction studies

The identification of disease specific proteins alone is insufficient to begin a drug screening process. It is critical to assign function and validation to these proteins by confirming they are indeed pivotal in the disease process. These studies need to encompass both gain- and loss-of-function analyses. This would determine whether the activity of a candidate target (an enzyme, for example), eliminated by molecular/cellular techniques, could reverse a disease phenotype. If this happened, then the investigator would have increased confidence that a small-molecule inhibitor against the target would also have a similar effect. The proposal of candidate drug targets is often not a difficult process, but validating them is another matter. Validation represents a major bottleneck where the wrong decision can have serious consequences[20].

Proteomics can be used to evaluate the role of a chosen target protein in signal transduction cascades directly relevant to the disease. In this manner, valuable information is forthcoming on the signalling pathways that are perturbed by a target protein and how they might be corrected by appropriate therapeutics. Techniques that are well established in one-dimensional protein studies to investigate signalling pathways, such as western blotting and immunoprecipitation, are highly suited to proteomic applications. For example, the proteomes obtained can be blotted onto membranes and probed with antibodies against the target protein or related signalling molecules[21–23]. Because proteomics can resolve >2000 proteins on a single gel, it is possible to derive important information on specific isoforms (such as glycosylated or phosphorylated variants) of signalling molecules. This will result in characterization of how they are altered in the disease process. Western immunoblotting techniques using high-affinity antibodies will typically identify proteins present at ~10 copies per cell (~1.7 fmol); this is in contrast to the best fluorescent dyes currently available that are limited to imaging proteins at 1000 or more copies per cell. The level of sensitivity derived by these applications will greatly facilitate interpretation of complex signalling pathways and contribute significantly to validation of the target under study.

Immunoprecipitation studies

Similarly, immunoprecipitation studies are another useful way to exploit the resolving power of proteomics[24,25]. In this instance, very large quantities of protein (e.g. several milligrams) can be subjected to incubation with antibodies against chosen signalling molecules. This allows high-affinity capture of these proteins, which can subsequently be eluted and electrophoresed on a 2D gel to provide a high-resolution proteome of a specific subset of proteins. Detection by blot analysis allows the identification of extremely small amounts of defined signalling molecules. Again, the different isoforms of even very low abundance proteins can be seen, and, very importantly, the technique allows the investigator to identify multiprotein complexes or other proteins that co-precipitate with the target protein. These coassociating proteins frequently represent signalling partners for the target protein, and their identification by mass spectrometry can lead to invaluable information on the signalling processes involved.

The depth of signal transduction analysis offered by proteomics, and the utility for target validation studies, can be extended even further by applying cell fractionation studies[26–28]. By purifying subcellular fractions, such as membrane, nuclear, organelle and cytosolic, it is possible to assign a localization to proteins of interest and to follow their trafficking in a cell. Enrichment of these fractions will also allow much higher representation of low abundance proteins on the proteome. Their detection by fluorescent dyes or immunoblot techniques will lead to the identification of proteins in the range of 1–10 copies per cell, putting the sensitivity on a par with genomic approaches.

These signal transduction analyses can be of additional value in experiments where inhibitors derived from a screening programme against the target are being evaluated for their potency and selectivity. The inhibitors can encompass small molecules, antisense nucleic acid constructs, dominant-negative proteins, or neutralizing antibodies microinjected into cells. In each case, proteome analysis can provide unique data in support of validation studies for a chosen candidate drug target.

Proteomics and drug mode-of-action studies

Once a validated target is committed to a screening regimen to identify and advance a lead molecule, it is important to confirm that the efficacy of the inhibitor is through the expected mechanism. Such mode-of-action studies are usually tackled by various cell biological and biochemical methods. Proteomics can also be usefully applied to these studies and this is illustrated below by describing data obtained with OGT719. This is a novel galactosyl derivative of the cytotoxic agent 5-fluorouracil (5-FU), which is currently being developed by OGS for the treatment of hepatocellular carcinoma and colorectal metastases localized in the liver. The premise underpinning the design and rationale of OGT719 was to derive a 5-FU prodrug capable

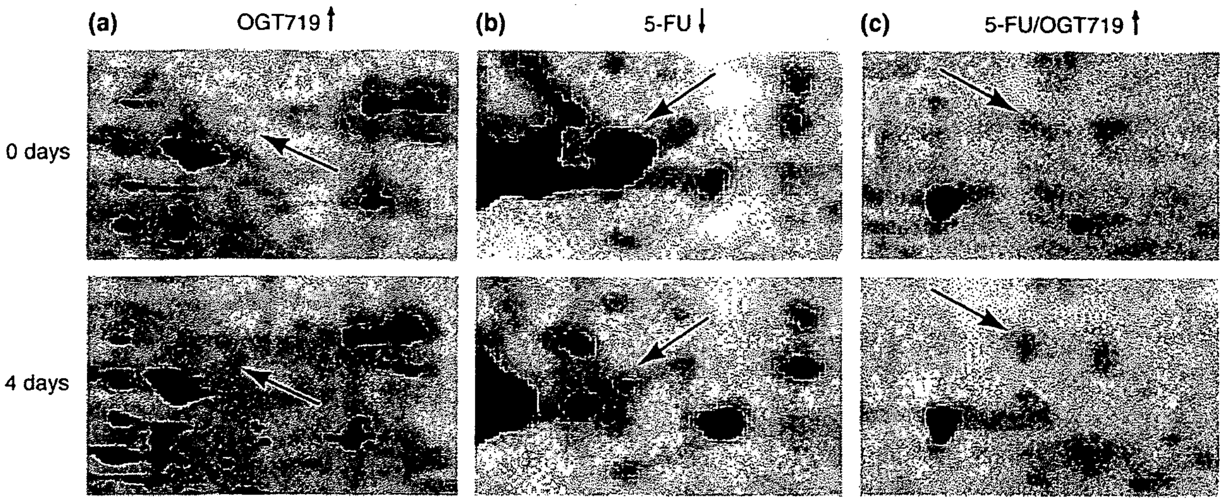


Figure 4. *Features that are specifically up- or downregulated in Huh7 cells by either 5-fluorouracil (5-FU) or OGT719: (a) elongation factor 1α2, (b) novel (three peptides by MS-MS) and (c) α-subunit of prolyl-4-hydroxylase. Arrows indicate up- or downregulated.*

of targeting, and being retained in, cells bearing the asialoglycoprotein receptor (ASGP-r), including hepatocytes[29], hepatoma Huh7 cells[30] and some colorectal tumour cells[31]. The growth of the human hepatoma cell line Huh7 is inhibited by 5-FU or by OGT719. If the inhibition by OGT719 were the result of uptake and conversion to 5-FU as the active component, then it would be expected that Huh7 cells would show similar proteome profiles following exposure to either drug.

To examine these possibilities, we conducted an experiment taking samples of Huh7 cells that had been treated with IC_{50} doses of either OGT719 or 5-FU. Total cell lysates were prepared and taken through 2D electrophoresis, fluorescence staining, digital imaging and Proteograph analysis. To facilitate the interpretation of the data across all of the 2291 features seen on the proteomes, drug-induced protein changes of fivefold or greater, identified by the Proteograph, were analysed further. Interestingly, from this analysis 19 identical proteins were changed fivefold or more by both drugs, strongly suggesting similarities in the mode of action for these two compounds.

Thus, from very complex data involving >2000 protein features, using proteomics it is possible to analyse quantitatively and qualitatively each protein during its exposure to drugs. The biologist is now able to focus a series of further studies specifically on an enriched subset of proteins. Figure 4 shows highlighted examples of the selected areas of the proteome where some of these identified proteins in the above study are altered in response to either or both drugs.

Several of the proteins identified above as being modulated similarly by 5-FU or OGT719 in Huh7 cells were subjected to tandem mass-spectrometric analysis for annotation. Some of these, such as the nuclear ribosomal RNA-binding protein[32], can be placed into pyrimidine pathways or related cell cycle/growth biochemical pathways in which 5-FU is known to act.

To attribute further significance to the proteome mode-of-action studies with OGT719, another cell line, the rat sarcoma HSN, was used. Growth of these cells is inhibited by 5-FU, but they are completely refractory to OGT719; notably they lack the ASGP-r, which might explain this finding (unpublished). For our proteome studies, HSN cells were treated with 5-FU or OGT719 over a time course of one, two and four days. At each time point, cells were harvested and processed to derive proteomes and Proteographs. As before, we purposely focused on those proteins that increased or decreased by fivefold or more. In this instance, there were no proteins co-modulated by the two drugs. This is perhaps to be expected, given that the HSN cells are killed by 5-FU and yet are refractory to OGT719.

Clear potential

The above is just an example of how proteomics can be used to address the mode of action of anticancer drugs. The potential of this approach is clear, and one can envisage situations where it will be profitable to compare the proteomes of cells in which the drug target has been eliminated by molecular knockout techniques, or with small-molecule inhibitors believed to act specifically on the same target. In addition to using proteomics to examine the action of drugs, it is also possible to use this approach to gauge the extent of nonspecific effects that might eventually lead to toxicity. For instance, in the example used above with HSN cells treated with OGT719, although cell growth was not affected, the levels of several specific proteins were changed. Further investigation of these proteins and the signalling pathways in which they are involved could be illuminating in predicting the likelihood or otherwise of long-term toxicity.

Use of proteomics in formal drug toxicology studies

A drug discovery programme at the stage where leads have been identified and mode-of-action studies are advanced, will proceed to investigate the pharmacokinetic and toxicology profile of those agents. These two parameters are of major importance in the drug discovery process, and many agents that have looked highly promising from *in vitro* studies have subsequently failed because of insurmountable pharmacokinetic and/or toxicity problems *in vivo*. Whereas the pharmacokinetic properties of a molecule can now be characterized quickly and accurately, toxicity studies are typically much longer and more demanding in their interpretation.

The ability to achieve fast and accurate predictions of toxicity within an *in vivo* setting would represent a big step forward in accelerating any drug discovery programme. Toxicity from a drug can be manifested in any organ. However, because the liver and kidney are the major sites in the body responsible for metabolism and elimination of most drugs, it is informative to examine these particular organs in detail to provide early indications about events that might result in toxicity.

The basis for most xenobiotic metabolizing activity is to increase the hydrophilicity of the compound and so facilitate its removal from the body. Most drugs are metabolized in the liver via the cytochrome P450 family of enzymes, which are known to comprise a total of ~200 different members[33,34], encompassing a wide array of overlapping specificities for different substrates. In addition to clearance, they also play a major role in metabolism that can lead to the production and removal of toxic species, and in some instances it is possible to correlate the ability or failure to remove such a toxin with a specific P450 or subgroup.

Unique P450 profiles

Each individual person will have a slightly different P450 profile, largely from polymorphisms and changes in expression levels, although other genetic and environmental factors aside from P450 also need to be taken into consideration. A significant amount of research is currently being directed towards this field – known as pharmacogenomics – with the aim of predicting how a patient will respond to a drug, as determined by their genetic makeup[35–37]. The marked variation of individuals in their ability to clear a compound can be one of the key factors in deciding the overall pharmacokinetic profile of a drug. Not only will this have a bearing on the likelihood of a patient responding to a treatment, but it will also be a factor in determining the possibility of their experiencing an adverse effect.

Many pharmaceutical companies are already employing genomic approaches, involving P450 measurements, as a key step in their assessment of the toxicological profile of a candidate drug and therefore of its suitability, or otherwise, to be considered for human clinical trials. There are limits to this approach, however. Whereas the P450 mRNA profiling can predict with some accuracy the likely metabolic fate of a drug, it will not provide information on whether the metabolites would subsequently lead to toxicity. Besides the patient-to-patient differences in steady-state levels of the P450s, there are also characteristic induction responses of these enzymes to some drugs. Moreover, as there can be some doubt over the correlation of mRNA levels and the corresponding protein levels, there is scope for misinterpretation of the results and hence real advantages to be gained from a proteome approach. In both instances, the ability to examine entire proteome profiles, including the P450 proteins, will be a significant advantage in understanding and predicting the metabolism and toxicological outcome of drugs.

In addition to direct organ and tissue studies, the serum, which collects the majority of toxicity markers released from susceptible organs and tissues throughout the entire body, can be utilized. Serum is rich in nuclease activity and, as pharmacogenomics is not suited to deal with these samples, valuable markers of toxicity could go undetected. However, by using proteomics for these types of analyses, serum markers (and clusters thereof) are now accessible for evaluation as indicators of toxicity.

Pharmacoproteomics

Proteomics can thus be used to add a new sphere of analysis to the study of toxicity at the protein level, and in the era of '-omics' there is a case to be made to adopt the term 'Pharmacoproteomics™'. Animals can be dosed with increasing levels of an experimental drug over time, and serum samples can be drawn for consecutive proteome analyses. Using this procedure, it should be possible to identify individual markers, or clusters thereof, that are dose related and correlate with the emergence and severity of toxicity. Markers might appear in the serum at a defined drug dose and time that are predictive of early toxicity within certain organs and if allowed to continue will have damaging consequences. These serum markers could subsequently be used to predict the response of each individual and allow tailoring of therapy whereby optimal efficacy is achieved without adverse side effects being apparent. This application can obviously extend to tracking toxicity of drugs in clinical trials where serum can be readily drawn and analysed. Surrogate markers for drug efficacy could also be detected by this procedure and could facilitate the challenge of identifying patient classes who will respond favourably to a drug and at what dosage.

Conclusions

By contrast to the agents administered to patients in clinical wards, the process of drug discovery is not a prescriptive series of steps. The risks are high and there are long timelines to be endured before it is known whether a candidate drug will succeed or fail. At each step of the drug discovery process there is often scope for flexibility in interpretation, which over many steps is cumulative. The pharmaceutical companies most likely to succeed in this environment are those that are able to make informed accurate decisions within an accelerated process.

The genomics revolution has impacted very positively upon these issues and now has a powerful new partner in proteomics. The ability to undertake global analysis of proteins from a very wide diversity of biological systems and to interrogate these in a high-throughput, systematic manner will add a significant new dimension to drug discovery. Each step of the process from target discovery to clinical trials is accessible to proteomics, often providing unique sets of data. Using the combination of genomics and proteomics, scientists can now see every dimension of their biological focus, from genes, mRNA, proteins and their subcellular localization. This will greatly assist our understanding of the fundamental mechanistic basis of human disease and allow new improved and speedier drug discovery strategies to be implemented.

REFERENCES

1 Crooke, S.T. (1998) *Nat. Biotechnol.* 16, 29–30
2 Dykes, C.W. (1996) *Br. J. Clin. Pharmacol.* 42, 683–695
3 Schena, M. *et al.* (1998) *Trends Biotechnol.* 16, 301–306
4 Ramsay, G. (1998) *Nat. Biotechnol.* 16, 40–44
5 Anderson, N.L. and Anderson, N.G. (1998) *Electrophoresis* 19, 1853–1861
6 James, P. (1997) *Biochem. Biophys. Res. Commun.* 231, 1–6
7 Wilkins, M.R. *et al.* (1996) *Biotechnol. Genet. Eng. Rev.* 13, 19–50
8 Parekh, R.B. and Rohlff, C. (1997) *Curr. Opin. Biotechnol.* 8, 718–723
9 Figeys, D. *et al.* (1998) *Electrophoresis* 19, 1811–1818
10 Wimmer, K. *et al.* (1996) *Electrophoresis* 17, 1741–1751
11 Giometti, C.S., Williams, K. and Tollaksen, S.L. (1997) *Electrophoresis* 18, 573–581
12 Williams, K. *et al.* (1998) *Electrophoresis* 19, 333–343
13 Rasmussen, R.K. *et al.* (1998) *Electrophoresis* 19, 818–825
14 Hirano, T. *et al.* (1995) *Br. J. Cancer* 72, 840–848
15 Ji, H. *et al.* (1997) *Electrophoresis* 18, 605–613
16 Ostergaard, M. *et al.* (1997) *Cancer Res.* 57, 4111–4117
17 Patel, V.B. *et al.* (1997) *Electrophoresis* 18, 2788–2794
18 Arnott, D. *et al.* (1998) *Anal. Biochem.* 258, 1–18
19 Anderson, L. and Seilhamer, J. (1997) *Electrophoresis* 18, 533–537
20 Rastan, S. and Beeley, L.J. (1997) *Curr. Opin. Genet. Dev.* 7, 777–783
21 Gravel, P. *et al.* (1995) *Electrophoresis* 16, 1152–1159
22 Qian, Y. *et al.* (1997) *Clin. Chem.* 43, 352–359
23 Sanchez, J.C. *et al.* (1997) *Electrophoresis* 18, 638–641
24 Watts, A.D. *et al.* (1997) *Electrophoresis* 18, 1086–1091
25 Asker, N. *et al.* (1995) *Biochem. J.* 308, 873–880
26 Ramsby, M.L., Makowski, G.S. and Khairallah, E.A. (1994) *Electrophoresis* 15, 265–277
27 Huber, L.A. (1995) *FEBS Lett.* 369, 122–125
28 Corthals, G.L. *et al.* (1997) *Electrophoresis* 18, 317–323
29 Hubbard, A.L., Wall, D.A. and Ma, A. (1983) *J. Cell Biol.* 96, 217–229
30 Zeng, F.Y., Oka, J.A. and Weigel, P.H. (1996) *Biochem. Biophys. Res. Commun.* 218, 325–330
31 Mu, J-Z. *et al.* (1994) *Biochim. Biophys. Acta* 1222, 483–491
32 Ghoshal, K. and Jacob, S.T. (1997) *Biochem. Pharmacol.* 53, 1569–1575
33 Guengerich, F.P. and Parikh, A. (1997) *Curr. Opin. Biotechnol.* 8, 623–628
34 Rendic, S. and Di Carlo, F.J. (1997) *Drug Metab. Rev.* 29, 413–580
35 Vermes, A., Guchelaar, H.J. and Koopmans, R.P. (1997) *Cancer Treat. Rev.* 23, 321–339
36 Housman, D. and Ledley, F.D. (1998) *Nat. Biotechnol.* 16, 492–493
37 Persidis, A. (1998) *Nat. Biotechnol.* 16, 209–210

Reference H
with Response dated 4/23/04
In USSN: 10/049,742

FOCUS - 17 of 19 DOCUMENTS

PR Newswire

August 11, 1997, Monday

SECTION: Financial News

DISTRIBUTION: TO BUSINESS AND MEDICAL EDITORS

LENGTH: 478 words

HEADLINE: Eli Lilly & Co. and **Acacia Biosciences** Enter Into Research Collaboration;
First Corporate Agreement for Acacia's Genome Reporter Matrix(TM)

DATELINE: RICHMOND, Calif., Aug. 11

BODY:

Acacia Biosciences and Eli Lilly and Company (Lilly) announced today the signing of a joint research collaboration to utilize Acacia's Genome Reporter Matrix(TM) (GRM) to aid in the selection and optimization of lead compounds. Under the collaboration, Acacia will provide chemical and biological profiles on a class of Lilly's compounds for an undisclosed fee.

Acacia's GRM is an assay-based computer modeling system that uses yeast as a miniature ecosystem. The GRM can profile the extent, nature and quantity of any changes in gene expression. Because of the similarities between the yeast and human genome, the system serves as an excellent surrogate for the human body, mimicking the effects induced by a biologically active molecule.

"Using yeast as a model organism for lead optimization makes a lot of sense given the high degree of homology with human metabolic pathways," said William Current of Lilly Research Laboratories. "Acacia's innovative GRM has the potential to provide enormous insight into the therapeutic impact of our compounds and make the drug discovery process more rational. It should substantially accelerate the development process."

"This first agreement with a major pharmaceutical company is an important milestone in the development of Acacia," said Bruce Cohen, President and CEO of Acacia. "The deal is in line with our strategy of establishing alliances that will allow our collaborators to use genomic profiles to identify and optimize compounds within their existing portfolios. In the long run, this technology can be used to characterize large scale combinatorial libraries, predict side effects prior to clinical trials and resurrect drugs that have failed during clinical trials."

The GRM incorporates two critical elements: chemical response profiles and genetic response profiles. The chemical response profiles measure the change in gene expression caused by potential therapeutics and then rank genes with altered expressions by degree of response. The genetic response profiles measure changes in gene expression caused by mutations in the genes encoding potential targets of pharmaceuticals; these genetic response profiles represent gold standards in drug discovery by defining the response profile expected for drugs with perfect selectivity and specificity. By comparing the two profiles, one can analyze a potential drug candidate's ability to mimic the action of a 'perfect' drug.

Acacia Biosciences is a functional genomics company developing proprietary technologies to enhance the speed and efficacy of drug discovery and development. Acacia's Genome Reporter Matrix capitalizes on the latest advances in genomics and combinatorial chemistry to generate comprehensive profiles of drug candidates' in vivo activity.
SOURCE **Acacia Biosciences**

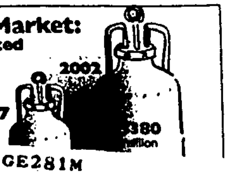
CONTACT: Bruce Cohen, President and CEO of **Acacia Biosciences**, 510-669-2330 ext. 103 or Media: Linda Seaton of Feinstein

LOAD-DATE: August 12, 1997

Reference I
with Response dated 4/23/04
In USSN: 10/049,742

The Bioreactor Market:
Steady Growth Expected

The worldwide market for all bioreactors was valued at $275 million for 1997, and is expected to be worth $380 million by 2002. Tis— —engineering, gene



GENETIC ENGINEERING NEWS
GEN
BIOTECHNOLO— • BIOPROCESS • BIORESEARCH • TECHNOLOGY TRANSFER

Contents

Pharmagene Raises More Capital for Research on Human Tissues

By Sophia Fox

Pharmagene, the Royston, U.K.-based biopharmaceutical company specialising in the use of human biomaterials for drug discovery research, has raised a further £5 million from a group of investors led by 3i and Abacus Nominees. The funding will enable the company to expand both its human biomaterials collection and its capabilities across a range of proprietary platform technologies.

Gordon Baxter, Ph.D., Pharmagene's cofounder and chief operating officer, claimed, "by the end of this year Pharmagene will have access to the largest collection of human RNAs and proteins anywhere in the world, and a range of innovative, yet robust technologies

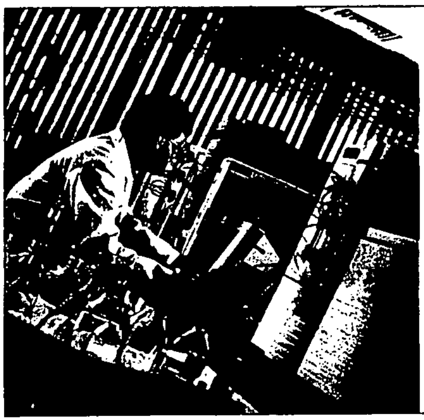
SEE PHARMAGENE, P. 9

Perkin-Elmer Acquires PerSeptive to Expand Its Capabilities in Gene-Based Drug Discovery

By John Sterling

Perkin-Elmer's (PE; Norwalk, CT) decision last month to acquire PerSeptive Biosystems (Framingham, MA) via a $360 million stock swap was designed to strengthen PE in terms of broad capabilities in gene-based drug discovery. The company's main goal is to develop new products to improve the integration of genetic and protein research.

"This merger will enhance our position as an effective provider of innovative, integrated platforms enabling our customers to be more efficient and cost-effective in bringing new pharmaceuticals to market," says Tony L. White, PE's chairman, president and CEO. "The combination of our two companies should bolster our presence in the life sciences, [and it is our] belief that we must take bold action now to lead the emerging era of molecular medicine with leading positions in both genetic and protein analysis."



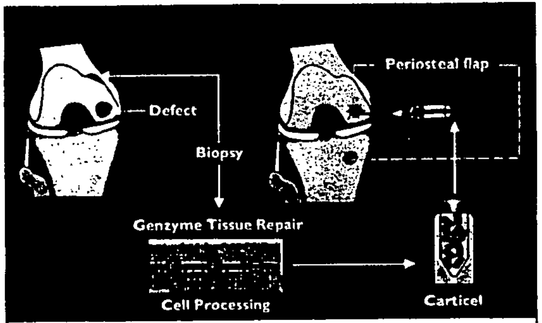
Perkin-Elmer acquired PerSeptive Biosystems for $360 million to obtain new technologies in mass spectrometry, bioseparations and purification for product development projects, spanning the range from genomics to proteomics.

A driving force behind the merger is the vast amount of genetic information about human disease that is being accumulated by researchers and biotech companies working in the area of genomics. It is becoming increasingly obvious that these data need to be complemented with technologies for studying proteins and protein networks—a field known as proteomics (*see* GEN, *September 1, 1997, p.1*).

PE officials, who claim that MALDI-TOF (Matrix Assisted

SEE ACQUISITION, P. 10

FDA OKs Genzyme's Carticel Product for Damage to Knees



Carticel, which was approved for the repair of clinically significant, symptomatic cartilaginous defects of the femoral condyle (medial, lateral or trochlear) caused by acute or repetitive trauma, employs a proprietary process to grow autologous cartilage cells for implantation.

By Naomi Pfeiffer

The FDA has approved a knee-cartilage replacement product made by Genzyme Tissue Repair (Cambridge, MA), a tracking-stock division of Genzyme Corp., for people with trauma-damaged knees.

Carticel™ (autologous cultured chondrocytes) is the first product to be licensed under the FDA's pro-

SEE GENZYME, P. 6

Strategies for Target Validation Streamline Evaluation of Leads

By Vicki Glaser

Acacia Biosciences (Richmond, CA) last month announced its first agreement with a major pharmaceutical company, signing a deal with Eli Lilly (Indianapolis, IN) to use Acacia's Genome Reporter Matrix (GRM) to select and optimize some of Lilly's lead compounds. Acacia's yeast-based system for profiling drug activity is useful for evaluating the therapeutic potential of lead compounds, and it also has a role in the identification and validation of new drug targets.

"We're using the ecosystem of a cell to allow us to deduce the mechanism of action and target for any chemical," explains Bruce Cohen, president and CEO. "We screen for every target in a cell simultaneously...using transcription as a readout for how a cell is adapting to any perturbation," he says.

The GRM technology consists of two main databases: one is the genetic response profile, showing the effects of mutations in each individual yeast gene and compensatory gene regulatory mechanisms; the other is the chemical response profile, which documents changes in gene expression in response to chemical compounds. Computational analysis and pattern matching between the genetic and chemical profiles yields information on the specificity, potency and side-effects risk of a drug lead.

Targeting Targets

No longer is mapping and sequencing a gene—or the human genome—an end unto itself, but

SEE TARGET, P. 15

Sticky Ends

Avigen received two grants from the NIH & University of California for research on gene therapy for treatment of cancer & HIV infections...**MRL Pharmaceutical Services**, of Reston, VA, launched the TSN Bug Finder, which is able to locate & retrieve client-specified microorganisms in real-time...**Gensia Sicor, Inc.** will move its corporate staff from San Diego to Irvine, CA, by end of year...FDA accepted NDA from **Sepracor** for levalbuterol HCl inhalation solution...An $11.7M mezzanine financing has been closed by **Activated Cell Therapy**, which changed its name to **Dendreon Corporation**...**Astra AB** will build major research facility in Waltham, MA, and is also relocating **Astra Arcus** research facility from Rochester to Boston area...**Prolifix Ltd.** team used a small peptide to inhibit the E2F protein complex and induced apoptosis in mammalian tumor cells...**Vertex Pharmaceuticals, Inc.** and **Alpha Therapeutic Corp.** ended an agreement to develop VX-366 for treatment of inherited hemoglobin disorders...**NaviCyte** received Phase I SBIR grant for up to $100,000 from NIH for development of prototype of its NaviFlow technology for high-throughput screening...**Covance Inc.** will invest $21 million in expansion and renovation of its facility in Indianapolis, IN.

PUBLISHED BY Mary Ann Liebert, Inc. publishers • NEW YORK

Target

from page 1

merely a means to an end. The critical next step is to validate the gene and its protein product as a potential drug target. The Human Genome Project continues to produce a treasure chest of expressed sequence tags (ESTs) and a tantalizing array of complete gene sequences.

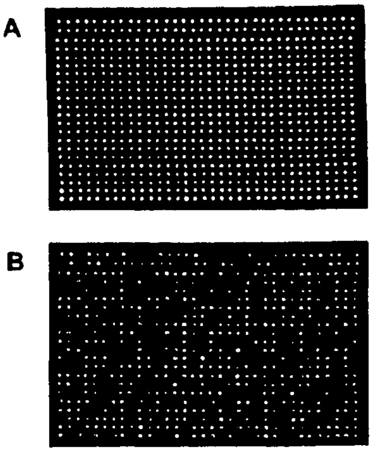
Companies are applying a variety of functional genomic strategies to link genes to specific diseases and to multigenic phenotypes. Yet the ultimate challenge for pharmaceutical companies is to sift through all the sequence and differential gene expression data to identify the best targets for drug discovery.

Spinning off technology developed at the University of North Carolina (Chapel Hill), **Cytogen Corp.** (Princeton, NJ) formed its wholly owned subsidiary **AxCell Biosciences** earlier this year. The young company is building a protein interaction database, cataloging all the interactions the modular domains of proteins can engage in with a range of ligands, in order to gain insight into protein function and to select the most critical interaction to target for drug development.

AxCell's cloning-of-ligand-targets (COLT) technology employs "recognition units" from the company's genetic diversity library (GDL) to map functional protein interactions and quantitate their affinity. The company's inter-functional proteomic database (IFP-dbase) elucidates protein interaction networks and structure-activity relationships based on ligand affinity with protein modular domains.

Defining Disease Pathways

Signal Pharmaceuticals, Inc.'s (San Diego, CA) integrated drug target and discovery effort is based on mapping gene-regulating pathways in cells and identifying small molecules that regulate the activation of those genes. In collaboration with academic researchers, the company has identified a large number of regulatory proteins in several mitogen-activated protein (MAP) kinase pathways (including the JNK, FRK and p38 signaling pathways), which Signal is evaluating for the treatment of autoimmune, inflammatory, cardiovascular and neurologic diseases, and cancer. Other target identification programs focus on the NF-kB pathway, estrogen-related genes and central/peripheral nervous system genes.



The Genome Reporter Matrix depicts a subset of a yeast array. Each colony harbors a GFP-reporter construct for a single gene. Collectively, the array reports the expression of all yeast genes.
A: Array in visible light.
B: Image of fluorescent emission from the array.

Acacia Biosciences

Regulating cytokine production in immune and inflammatory disorders, and modifying bone metabolism to treat osteoporosis are the focus of Signal's collaboration with **Tanabe Seiyaku** (Osaka, Japan). Signal has partnered with **Organon/Akzo Nobel** (Netherlands) to identify estrogen-responsive genes as targets for treating neurodegenerative and psychiatric diseases, atherosclerosis and ischemia, and with **Roche Bioscience** (Palo Alto, CA) to develop human peripheral nerve cell lines for the discovery of treatments for pain and incontinence.

Exelixis' (S. San Francisco, CA) strategy for target selection is to define disease pathways and identify regulatory molecules that activate or inhibit those biochemical/genetic pathways. Based on the finding that these pathways are conserved across species, the company is studying the model genetic systems of Drosophila and *Caenorhabditis elegans*. Using its PathFinder technology, Exelixis systematically introduces mutations into the genomes of these model organisms, looking for mutations that enhance or suppress the target disease-related gene. These novel genes then become the basis of drug screening assays.

Cadus Pharmaceutical Corp. (Tarrytown, NY) is identifying surrogate ligands to newly discovered orphan G-protein coupled transmembrane receptors of unknown function to determine the suitability of the receptors as drug targets. Inserting the novel receptor in a yeast system yields a ligand that activates the receptor. Access to a surrogate ligand allows the company to screen for receptor antagonists in the yeast system.

"The antagonist plus the surrogate ligand gives you two probes—an on probe and an off probe—which allows you to look at function," explains David Webb, Ph.D., vp of research and chief scientific officer. A surrogate ligand also provides information on which G-protein interacts with the orphan receptor and its associated signaling pathways, further clarifying the role of the receptor as a potential drug target. Cadus' collaboration with **SmithKline** (Philadelphia) capitalizes on Cadus' ability to determine orphan receptor function, applying the technology to SmithKline's proprietary, newly discovered G-protein receptors.

Cadus' recombinant yeast system can also be used to screen cell and tissue extracts for natural ligands, and the company is accelerating its internal drug-discovery efforts in the areas of cancer, inflammation and allergy. A recent equity investment in **Axiom Biotechnologies** (San Diego, CA) gave Cadus a license to Axiom's high-throughput pharmacologic screening system for lead optimization and discovery.

As its name implies, **gene/Networks** (Alameda, CA) focuses on identifying gene networks that contribute to multigenic phenotypes and complex disease processes. The integration of mouse and human genetic studies forms the basis of the technology. The Genome Tagged Mice database in development will serve as a library of natural mouse genetic and phenotypic variation. Disease-related genes identified in mice are then evaluated in human family- and population-based studies to confirm their clinical relevance and linkages to pathophysiologic traits.

Blocking Gene Expression

Inactivating a gene known to be expressed in association with a particular disease is one approach to identifying appropriate therapeutic targets. The target validation and discovery program at **Ribozyme Pharmaceuticals, Inc.** (Boulder, CO) applies the company's ribozyme technology to achieve selective inhibition of gene expression in cell culture and in animals.

Correlation of the gene expression inhibition with phenotype can

SEE TARGET, P. 38

Target

from page 15

suggest the relative importance of the gene in disease pathology. The company's nuclease-resistant ribozymes form the basis of a collaboration with **Schering AG** (Germany) for drug target validation and the development of ribozyme-based therapeutic agents, and with **Chiron Corp.** (Emeryville, CA) for target validation.

With several antisense compounds now progressing through clinical trials, the concept of using oligonucleotides to inhibit gene activity is not new. But rather than focusing on therapeutics development, **Sequitur, Inc.** (Natick, MA) is creating antisense compounds for the purpose of determining gene function and validating drug targets. Clients typically provide the one-year-old company with the sequence (or EST) of a potential gene target and, in return, Sequitur custom designs a series of three to six antisense compounds that yield a three-to-ten-fold inhibition of the target gene in cell culture. The company also provides oligofectins, a series of cationic lipids, to deliver the oligonucleotides to a variety of cultured cells.

"Differential expression information is just for correlation, it doesn't tell function or confirm what would be a good target," says Tod Woolf, Ph.D., director of technology development at Sequitur. Whereas, antisense compounds will inhibit a target, Sequitur offers both phosphorothioate DNA antisense compounds, and its proprietary Next Generation chimeric oligonucleotides, which have a higher hybridization affinity, greater specificity and reduced toxicity, according to the company.

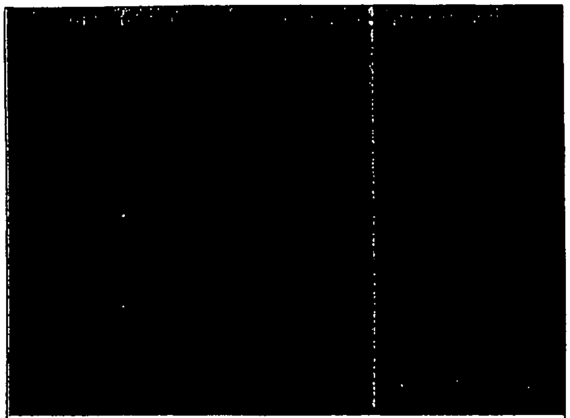
Mining Pathogen Genomes

Companies such as **Human Genome Sciences** (HGS; Rockville, MD), **Incyte** (Palo Alto, CA), **Millennium Pharmaceuticals Inc.** (Cambridge, MA) and **Genome Therapeutics** (Waltham, MA) are relying on high-speed DNA sequencing, positional cloning and other strategies to identify specific microbial genomic sites that would be good targets for infectious disease therapeutics.

HGS recently completed sequencing of the bacterial pathogen *Streptococcus pneumoniae*, which is the focus of an agreement with **Hoffmann-La Roche** (Basel, Switzerland). Roche will use the sequence data to develop new anti-infectives against *S. pneumoniae*. HGS and Roche have expanded their collaboration to include a nonexclusive license to access sequence information for the intestinal bacterium *Enterococcus faecalis*.

Incyte Pharmaceuticals has completed one-fold coverage of the *Candida albicans* genome, identifying 60% of the genes of this fungal pathogen. This genome will become part of the company's PathoSeq microbial database. Incyte recently introduced the ZooSeq animal gene sequence and expression database. The database will provide genomic information across various species commonly used in preclinical drug testing, which may help to better define potential drug targets.

Millennium Pharmaceuticals continues to report success in identifying novel drug targets, having recently discovered a novel chemokine called neurotactin and a new class of MAD-related proteins that inhibit transforming growth factor beta (TGF-ß) signaling. The company also received U.S. patent coverage for the tub genes, believed to play a role in obesity, and for the gene that encodes the protein melastatin, which appears to suppress metastasis in malignant melanoma. ■



AxCell Biosciences scientists say their technology enables the rapid and simple functional identification of the two essential molecular components of protein interaction networks: specific recognition units that bind distinct modular protein domains are identified and isoluted using a combination structural/functional approach that uses both peptide phase display Genetic Diversity Libraries (GDL) and bioinformatics, and cloning of Ligand Targets (COLT) technology utilizes recognition units as functional probes to isolate families of interactor proteins.

Pangea

from page 28

Smith, now a computer programmer, is an expert in systems integration, Internet technologies and the application of industrial engineering principles to the drug discovery process. Before co-founding Pangea, he was the manager of software development at Attorney's Briefcase, a legal research software company.

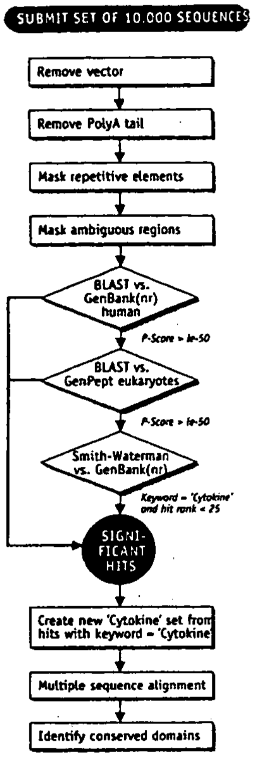
By being "in the trenches" with customers and collaborators, Bellenson and Smith sensed the frustration of pharmaceutical researchers whose incompatible tools have impeded their progress. According to Bellenson, "Most of them are geared toward analyzing one molecule at a time. It's like emptying the ocean with an eye dropper—an incompatible eye dropper at that. A pharmaceutical company may have 30 different drug discovery teams with various approaches. The problem is to manage the process of experimenting with a lot of different approaches, to automate while maintaining flexibility."

GeneWorld 2.1 enables "integration of the entire target discovery and validation process," Bellenson says. The commercial software package coordinates the entire process of sequence-data analysis and can be integrated with other programs and databases, according to Smith, who adds that it handles thousands of sequence results, organizes and automates annotation and seamlessly interacts with growing genome databases. Simple forms and menus enable users to turn raw sequence data into crucial knowledge for drug discovery by applying algorithms to sequences, creating custom analysis strategies and producing useful reports, without the need for writing computer code. GeneWorld 2.1 runs on a variety of platforms and operating systems.

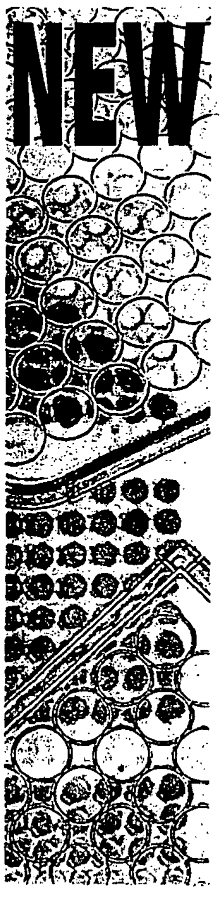
Pairing industrial relational database-management systems with a web-browser interface, Pangea's Operating System of Drug Discovery™ is an open-computing framework that allows client/server and Java-enabled web-based technologies to collect, organize and analyze drug discovery information for pharmaceutical companies to simplify and accelerate drug discovery. The technology unites automated genomics database analysis for drug target site selection, chemical information database analysis and large-scale combinatorial chemistry project management and high-throughput screening project management for drug lead efficacy analysis. Pangea officials maintain that these integrated elements provide a unified environment for chemists, biologists and others involved in the drug discovery process to work together with commercial and public domain software.

Pangea's Operating System of Drug Discovery can accommodate Sybase, Oracle or Informix relational database-management systems and any version of UNIX. It absorbs new data formats, databases, algorithms and analysis paradigms into the automated workflow without software modifications. Netscape Navigator™ provides a friendly user interface from PC, Macintosh, and UNIX workstations.

In the near term, Pangea plans to complete its bioinformatics core with two more programs. Gene Foundry, a sample tracking and workflow sequence package for DNA sequence and fragment information, will also offer interaction with robots, reagent tracking and troubleshooting. Gene Thesaurus, the other package is a "warehouse of bioinformatics data," says Bellenson. ■


SUBMIT SET OF 10.000 SEQUENCES
Remove vector
Remove PolyA tail
Mask repetitive elements
Mask ambiguous regions
BLAST vs. GenBank(nr) human
P-Score > le-50
BLAST vs. GenPept eukaryotes
P-Score > le-50
Smith-Waterman vs. GenBank(nr)
Keyword = 'Cytokine' and hit rank < 25
SIGNI-FICANT HITS
Create new 'Cytokine' set from hits with keyword = 'Cytokine'
Multiple sequence alignment
Identify conserved domains

Bioinformaticists can design and save Strategies, such as the one shown here, that forward data through multiple-step analyses logically and automatically. Researchers throughout your organization can apply the same Strategies to their own data.



Europe

from page 30

GTAC Chairman, Professor Norman C. Nevin, said 1996 saw "four important developments": an increase in enquiries and submissions made to GTAC; an increase in the complexity of submitted protocols; a continuing shift from gene therapy for single-gene disorders toward strategies aimed at tumour destruction in cancer; and a growth in international sponsorship of U.K. gene therapy trials.

Since 1993, GTAC and its predecessor, the Clothier Committee, have approved 18 U.K. gene therapy clinical trials (13 of which have been carried out), which are listed in the report. The disease areas targeted by these trials include severe combined immunodeficiency (1 trial), cystic fibrosis (6), metastatic melanoma (2), lymphoma (2), neuroblastoma (1), breast cancer (1), Hurler's syndrome (1), cervical cancer (1), glioblastoma breast cancer, breast cancer with liver metastases, glioblastoma, malignant ascites due to gastrointestinal cancer and ovarian cancer.

Copies of the GTAC thrid annual report are available from the GTAC Secretariat, Wellington House, 133-155 Waterloo Road, London SE1 8UG, U.K.

Coated Lenses Prevent PCO

Scientists in the U.K. say it may be possible to prevent posterior capsule opacification (PCO), a common complication following cataract surgery, by using the implanted polymethylmethacrylate (PMMA) intraocular lens as a drug delivery system. PCO occurs in 30-50% of cataract surgery patients as a result of stimulated cell growth within the remaining capsular bag. The condition causes a decline in visual acuity and requires expensive laser treatment, thus negating the routine use of cataract surgery in underdeveloped countries, explains G. Duncan, at the

Fischer-Vize, *Science* **270**, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* **6**, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 263 (1991); B. Tschiersch *et al.*, *EMBO J.* **13**, 3822 (1994); M. T. Madireddi *et al.*, *Cell* **87**, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 7137 (1996).
36. P. M. Palosaari *et al.*, *J. Biol. Chem.* **266**, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E. Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* **58**, 4068 (1992); V. Sharma, K. Suvarna, R. Meganathan, M. E. Hudspeth, *J. Bacteriol.* **174**, 5057 (1992); M. Kanazawa *et al.*, *Enzyme Protein* **47**, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* **178**, 3015 (1996).
37. M. Ho *et al.*, *Cell* **77**, 869 (1994).
38. W. Hendriks *et al.*, *J. Cell Biochem.* **59**, 418 (1995).

39. We thank H. Skaletsky and F. Lewitter for help with sequence analysis; Lawrence Livermore National Laboratory for the flow-sorted Y cosmid library; and P. Bain, A. Bortvin, A. de la Chapelle, G. Fink, K. Jegalian, T. Kawaguchi, E. Lander, H. Lodish, P. Matsudaira, D. Menke, U. RajBhandary, R. Reijo, S. Rozen, A. Schwartz, C. Sun, and C. Tilford for comments on the manuscript. Supported by NIH.

28 April 1997; accepted 9 September 1997

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (*1*, *2*), provide a practical and economical tool for studying gene expression on a very large scale (*3*–*6*).

Saccharomyces cerevisiae is an especially favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, *cis* regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (*7*). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (*8*). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by using a simple robotic printing device (*9*). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (*10*). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (*11*) and then hybridized to the microarrays (*12*). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (*13*).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (*14*). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305-5428, USA.

*To whom correspondence should be addressed. E-mail: brown@cmgm.stanford.edu

to any gene whose function is known (*15*). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (*ALD2*) and acetyl–coenzyme A(CoA) synthase (*ACS1*), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes *PCK1*, encoding phosphoenolpyruvate carboxykinase, and *FBP1*, encoding fructose 1,6-biphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c–related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (*13*). A noteworthy and illuminating exception was that the genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitchondrial biogenesis (*13*). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (*16–20*). A search in the promoter regions of the remaining two genes, *ACR1* and *IDP2*, revealed that *ACR1*, a gene essential for *ACS1* activity, also possessed a consensus CSRE motif, but interestingly, *IDP2* did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception

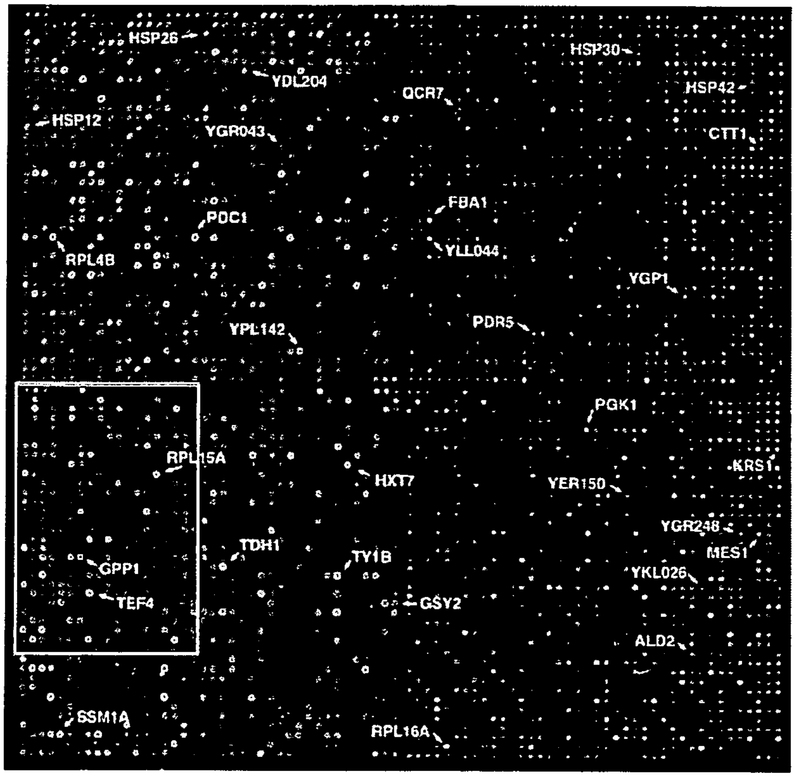


Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (*9*). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (*49*). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of <5 × 10^6 cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of ~2 × 10^8 cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP–labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP–labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of *HSP42*, have previously been shown to be controlled at least in part by these elements (*21–24*). Inspection of the sequences upstream of *HSP42* and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including *HSP30*, *ALD2*, *OM45*, and 10 uncharacterized ORFs (*25*)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex *HAP2,3,4* has been shown to be responsible for induction of several genes important for respiration (*26–28*). This complex binds a degenerate consensus sequence known as the CCAAT box (*26*). Computer analysis, using the consensus sequence TNRYTGGB (*29*), has suggested that a large number of genes involved in respiration may be specific targets of *HAP2,3,4* (*30*). Indeed, a putative *HAP2,3,4* binding site could be found in the sequences upstream of each of the seven cytochrome c–related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome c–related genes that were induced, *HAP2,3,4* binding sites were present in all but one. Significantly, we found that transcription of *HAP4* itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS_{rpg}) that is recognized by the Rap1 DNA-binding protein (*31*, *32*). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (*33*). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (*34*). Indeed, we observed that the abundance of *RAP1* mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, *HAP4* and *SIP4*, were induced by a factor of more than threefold at the diauxic shift. *SIP4* encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the "master regulator" of glucose repression (*35*). The eightfold induction of *SIP4* upon depletion of glucose strongly suggests a role in the induction of downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (*36*). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expression ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (*37*, *38*). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

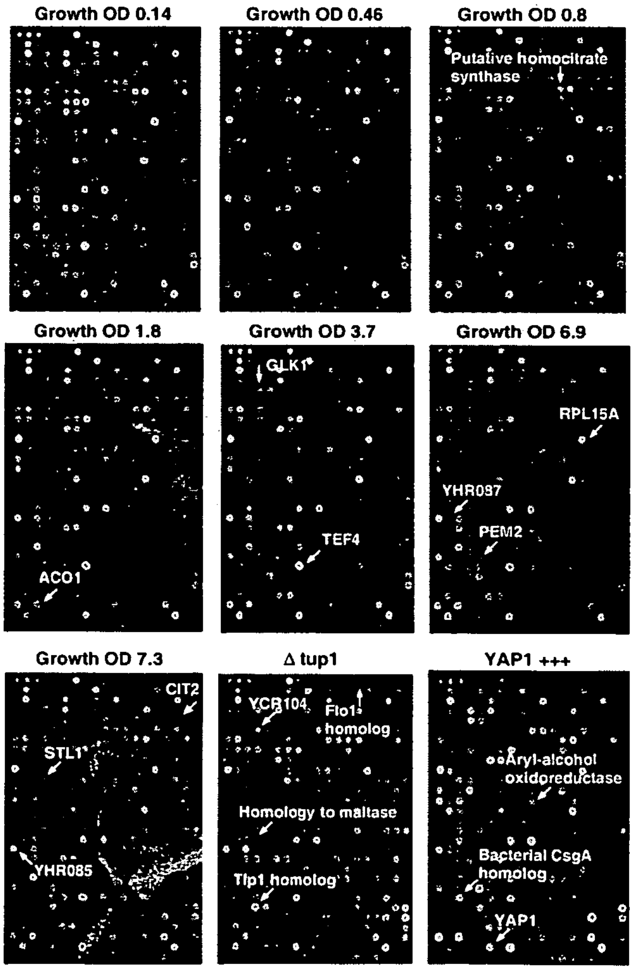


Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1Δ* mutation and *YAP1* overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (*13*). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.

by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (*39*). Tup1 has also been implicated in repression of oxygen-regulated, mating-type–specific, and DNA-damage–inducible genes (*40*).

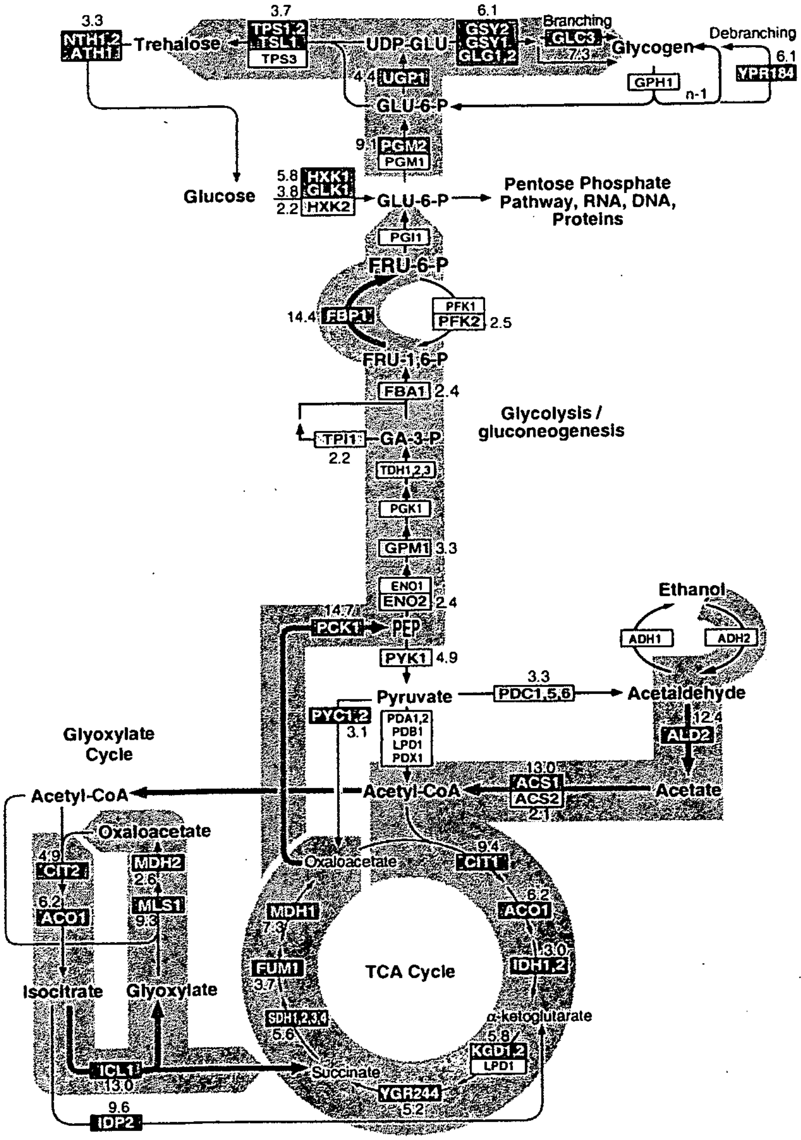


Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs YGR244 and YPR184 show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1*Δ) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (*11*). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1*Δ strain, and thus presumably repressed by Tup1 (*41*). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1*Δ mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (*13*)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α-glucosidases, the mating-type–specific genes *MFA1* and *MFA2*, and the DNA damage–inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1*Δ strain, providing a positive control in the experiment (*42*).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (*15*) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as Tip1 and Tir1/Srp1 which are induced by cold shock and other stresses (*43*), and similar, serine-poor proteins, the seripauperins (*44*). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1*Δ strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a *MAT* α strain in which *MFA1* and *MFA2*, the genes encoding the a-factor mating pheromone precursor, are normally repressed. In the isogenic *tup1*Δ strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a *MATA* strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, *o*-phenanthroline, heavy metals, and osmotic stress (*45*). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GAL1-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

Of the 17 genes whose mRNA levels increased by more than threefold when *YAP1* was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (*46*, *47*). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing Yap1. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for Yap1-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (*48*). About two-thirds of the genes that were induced by more than threefold upon Yap1 overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by Yap1. The absence of canonical Yap1-bind-

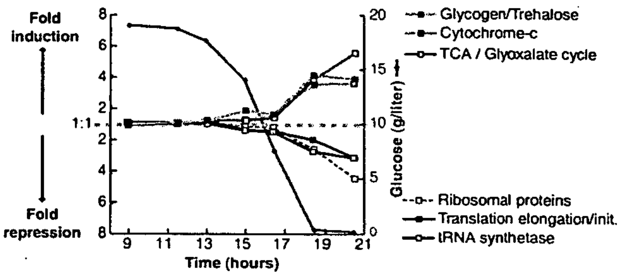


Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.

Table 1. Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (*50*). Positions of the canonical Yap1 binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of Yap1 site from ATG	Gene	Description	Fold-increase
YNL331C			Putative aryl-alcohol reductase	12.9
YKL071W	162–222 (5 sites)		Similarity to bacterial csgA protein	10.4
YML007W		YAP1	Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W			Putative aryl-alcohol reductase	6.5
YML116W	409	ATR1	Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	OYE3	NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	OYE2	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to *A. thaliana* zeta-crystallin homolog	3.7
YOL126C		MDH2	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and networks. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

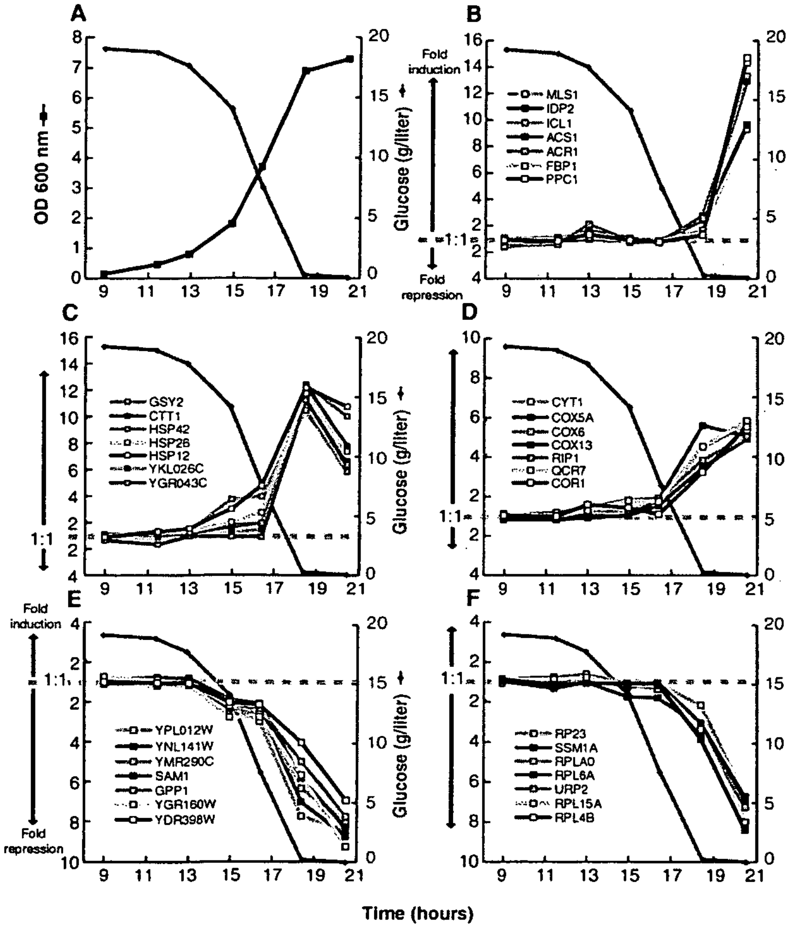


Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (**A**) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (**B**) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (**C**) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contain STRE motif repeats in their upstream promoter regions. (**D**) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (**E**) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (**F**) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* **6**, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi *et al.*, *Nature Genet.* **14**, 457 (1996).
5. D. J. Lockhart *et al.*, *Nature Biotechnol.* **14**, 1675 (1996).
6. M. Chee *et al.*, *Science* **274**, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast* Saccharomyces: *Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100-μl PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3× standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-

tion, the bound DNA was denatured by a 2-min incubation in distilled water at ~95°C. The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.

10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251) Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at −80°C.
11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 μg of polyadenylated [poly(A)$^+$] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 μM for dATP, dCTP, and dGTP and 200 μM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 μM. The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with of 470 μl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ~5 μl, using Centricon-30 microconcentrators (Amicon).
12. Purified, labeled cDNA was resuspended in 11 μl of 3.5× SSC containing 10 μg poly(dA) and 0.3 μl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~8 to 12 hours in a water bath at 62°C. Before scanning, slides were washed in 2× SSC, 0.2% SDS for 5 min, and then 0.05× SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.
13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html
14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.
15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: *Saccharomyces* Genome Database (genome-www.stanford.edu), Yeast Protein Database (quest7.proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/mips/yeast/index.htmlx).
16. A. Scholer and H. J. Schuller, *Mol. Cell. Biol.* **14**, 3613 (1994).
17. S. Kratzer and H. J. Schuller, *Gene* **161**, 75 (1995).
18. R. J. Haselbeck and H. L. McAlister, *J. Biol. Chem.* **268**, 12116 (1993).
19. M. Fernandez, E. Fernandez, R. Rodicio, *Mol. Gen. Genet.* **242**, 727 (1994).
20. A. Hartig *et al.*, *Nucleic Acids Res.* **20**, 5677 (1992).
21. P. M. Martinez *et al.*, *EMBO J.* **15**, 2227 (1996).
22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, *Mol. Cell. Biol.* **15**, 6232 (1995).
23. H. Ruis and C. Schuller, *Bioessays* **17**, 959 (1995).
24. J. L. Parrou, M. A. Teste, J. Francois, *Microbiology* **143**, 1891 (1997).
25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.
26. S. L. Forsburg and L. Guarente, *Genes Dev.* **3**, 1166 (1989).
27. J. T. Olesen and L. Guarente, *ibid.* **4**, 1714 (1990).
28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, *Mol. Microbiol.* **13**, 119 (1994).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B–C, G, or T; N–G, A, T, or C; R–A or G; and Y–C or T.
30. C. Fondrat and A. Kalogeropoulos, *Comput. Appl. Biosci.* **12**, 363 (1996).
31. D. Shore, *Trends Genet.* **10**, 408 (1994).
32. R. J. Planta and H. A. Raue, *ibid.* **4**, 64 (1988).
33. The degenerate consensus sequence VYCYRNNCMNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (*30*), is WACAYCCRTACATYW, with up to three differences allowed.
34. S. F. Neuman, S. Bhattacharya, J. R. Broach, *Mol. Cell. Biol.* **15**, 3187 (1995).
35. P. Lesage, X. Yang, M. Carlson, *ibid.* **16**, 1921 (1996).
36. For example, we observed large inductions of the genes coding for *PCK1*, *FBP1* [Z. Yin *et al.*, *Mol. Microbiol.* **20**, 751 (1996)], the central glyoxylate cycle gene *ICL1* [A. Scholer and H. J. Schuller, *Curr. Genet.* **23**, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, *ACS1* [M. A. van den Berg *et al.*, *J. Biol. Chem.* **271**, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes *PYK1* and *PFK2* [P. A. Moore *et al.*, *Mol. Cell. Biol.* **11**, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase T *CTT1* [P. H. Bissinger *et al.*, *ibid.* **9**, 1309 (1989)] and several genes encoding small heat-shock proteins, such as *HSP12*, *HSP26*, and *HSP42* [I. Farkas *et al.*, *J. Biol. Chem.* **266**, 15602 (1991); U. M. Praekelt and P. A. Meacock, *Mol. Gen. Genet.* **223**, 97 (1990); D. Wotton *et al.*, *J. Biol. Chem.* **271**, 2717 (1996)].
37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, *FBP1* and *PCK1*) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (*46*).
38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, *ADH1* and *ADH2*, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as *HXK1/HXK2* (77% identical) [P. Herrero *et al.*, *Yeast* **11**, 137 (1995)], *MLS1/DAL7* (73% identical) (*20*), and *PGM1/PGM2* (72% identical) [D. Oh, J. E. Hopper, *Mol. Cell. Biol.* **10**, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.
39. F. E. Williams, U. Varanasi, R. J. Trumbly, *Mol. Cell. Biol.* **11**, 3307 (1991).
40. D. Tzamarias and K. Struhl, *Nature* **369**, 758 (1994).
41. Differences in mRNA levels between the tup1Δ and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the tup1Δ strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.
42. The tup1Δ mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of *TUP1* and an Eco R I site 124 base pairs before the stop codon of the *TUP1* gene.
43. L. R. Kowalski, K. Kondo, M. Inouye, *Mol. Microbiol.* **15**, 341 (1995).
44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, *Gene* **148**, 149 (1994).
45. D. Hirata, K. Yano, T. Miyakawa, *Mol. Gen. Genet.* **242**, 250 (1994).
46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, *Appl. Environ. Microbiol.* **60**, 1783 (1994).
47. A. Muheim *et al.*, *Eur. J. Biochem.* **195**, 369 (1991).
48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, *J. Biol. Chem.* **269**, 32592 (1994).
49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm.stanford.edu/pbrown. Images were scanned at a resolution of 20 μm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.
50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold)

51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the Tup1 deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997

Reference F
with Response dated 4/23/04
In USSN: 10/049,742

Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 2150–2155, March 1997
Biochemistry

Discovery and analysis of inflammatory disease-related genes using cDNA microarrays

(inflammation/human genome analysis/gene discovery)

Renu A. Heller*†, Mark Schena*, Andrew Chai*, Dari Shalon‡, Tod Bedilion‡, James Gilmore‡, David E. Woolley§, and Ronald W. Davis*

*Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305; ‡Synteni, Palo Alto, CA 94306; and §Department of Medicine, Manchester Royal Infirmary, Manchester, United Kingdom

Contributed by Ronald W. Davis, December 27, 1996

ABSTRACT cDNA microarray technology is used to profile complex diseases and discover novel disease-related genes. In inflammatory disease such as rheumatoid arthritis, expression patterns of diverse cell types contribute to the pathology. We have monitored gene expression in this disease state with a microarray of selected human genes of probable significance in inflammation as well as with genes expressed in peripheral human blood cells. Messenger RNA from cultured macrophages, chondrocyte cell lines, primary chondrocytes, and synoviocytes provided expression profiles for the selected cytokines, chemokines, DNA binding proteins, and matrix-degrading metalloproteinases. Comparisons between tissue samples of rheumatoid arthritis and inflammatory bowel disease verified the involvement of many genes and revealed novel participation of the cytokine interleukin 3, chemokine Groα and the metalloproteinase matrix metallo-elastase in both diseases. From the peripheral blood library, tissue inhibitor of metalloproteinase 1, ferritin light chain, and manganese superoxide dismutase genes were identified as expressed differentially in rheumatoid arthritis compared with inflammatory bowel disease. These results successfully demonstrate the use of the cDNA microarray system as a general approach for dissecting human diseases.

The recently described cDNA microarray or DNA-chip technology allows expression monitoring of hundreds and thousands of genes simultaneously and provides a format for identifying genes as well as changes in their activity (1, 2). Using this technology, two-color fluorescence patterns of differential gene expression in the root versus the shoot tissue of *Arabidopsis* were obtained in a specific array of 48 genes (1). In another study using a 1000 gene array from a human peripheral blood library, novel genes expressed by T cells were identified upon heat shock and protein kinase C activation (3).

The technology uses cDNA sequences or cDNA inserts of a library for PCR amplification that are arrayed on a glass slide with high speed robotics at a density of 1000 cDNA sequences per cm^2. These microarrays serve as gene targets for hybridization to cDNA probes prepared from RNA samples of cells or tissues. A two-color fluorescence labeling technique is used in the preparation of the cDNA probes such that a simultaneous hybridization but separate detection of signals provides the comparative analysis and the relative abundance of specific genes expressed (1, 2). Microarrays can be constructed from specific cDNA clones of interest, a cDNA library, or a select number of open reading frames from a genome sequencing database to allow a large-scale functional analysis of expressed sequences.

Because of the wide spectrum of genes and endogenous mediators involved, the microarray technology is well suited for analyzing chronic diseases. In rheumatoid arthritis (RA), inflammation of the joint is caused by the gene products of many different cell types present in the synovium and cartilage tissues plus those infiltrating from the circulating blood. The autoimmune and inflammatory nature of the disease is a cumulative result of genetic susceptibility factors and multiple responses, paracrine and autocrine in nature, from macrophages, T cells, plasma cells, neutrophils, synovial fibroblasts, chondrocytes, etc. Growth factors, inflammatory cytokines (4), and the chemokines (5) are the important mediators of this inflammatory process. The ensuing destruction of the cartilage and bone by the invading synovial tissue includes the actions of prostaglandins and leukotrienes (6), and the matrix degrading metalloproteinases (MMPs). The MMPs are an important class of Zn-dependent metallo-endoproteinases that can collectively degrade the proteoglycan and collagen components of the connective tissue matrix (7).

This paper presents a study in which the involvement of select classes of molecules in RA was examined. Also investigated were 1000 human genes randomly selected from a peripheral human blood cell library. Their differential and quantitative expression analysis in cells of the joint tissue, in diseased RA tissue and in inflammatory bowel disease (IBD) tissues was conducted to demonstrate the utility of the microarray method to analyze complex diseases by their pattern of gene expression. Such a survey provides insight not only into the underlying cause of the pathology, but also provides the opportunity to selectively target genes for disease intervention by appropriate drug development and gene therapies.

METHODS

Microarray Design, Development, and Preparation. Two approaches for the fabrication of cDNA microarrays were used in this study. In the first approach, known human genes of probable significance in RA were identified. Regions of the clones, preferably 1 kb in length, were selected by their proximity to the 3′ end of the cDNA and for areas of least identity to related and repetitive sequences. Primers were synthesized to amplify the target regions by standard PCR protocols (3). Products were

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

0027-8424/97/942150-6$2.00/0
PNAS is available online at **http://www.pnas.org.**

Abbreviations: RA, rheumatoid arthritis; MMP, matrix-degrading metalloproteinase; IBD, inflammatory bowel disease; LPS, lipopolysaccharide; PMA, phorbol 12-myristate 13-acetate; TNF-α, tumor necrosis factor α; IL, interleukin; TGF-β, transforming growth factor β; GCSF, granulocyte colony-stimulating factor; MIP, macrophage inflammatory protein; MIF, migration inhibitory factor; HME, human matrix metallo-elastase; RANTES, regulated upon activation, normal T cell expressed and secreted; Gel, gelatinase; VCAM, vascular cell adhesion molecule; ICE, IL-1 converting enzyme; PUMP, putative metalloproteinase; MnSOD, manganese superoxide dismutase; TIMP, tissue inhibitor of metalloproteinase; MCP, macrophage chemotactic protein.

†To whom reprint requests should be sent at the present address: Roche Bioscience, S3-1, 3401 Hillview Avenue, Palo Alto, CA 94304.

verified by gel electrophoresis and purified with Qiaquick 96-well purification kit (Qiagen, Chatsworth, CA), lyophilized (Savant), and resuspended in 5 μl of 3× standard saline citrate (SSC) buffer for arraying. In the second approach, the microarray containing the 1056 human genes from the peripheral blood lymphocyte library was prepared as described (3).

Tissue Specimens. Rheumatoid synovial tissue was obtained from patients with late stage classic RA undergoing remedial synovectomy or arthroplasty of the knee. Synovial tissue was separated from any associated connective tissue or fat. One gram of each synovial specimen was subjected to RNA extraction within 40 min of surgical excision, or explants were cultured in serum-free medium to examine any changes under *in vitro* conditions. For IBD, specimens of macroscopically inflamed lower intestinal mucosa were obtained from patients with Crohn disease undergoing remedial surgery. The hypertrophied mucosal tissue was separated from underlying connective tissue and extracted for RNA.

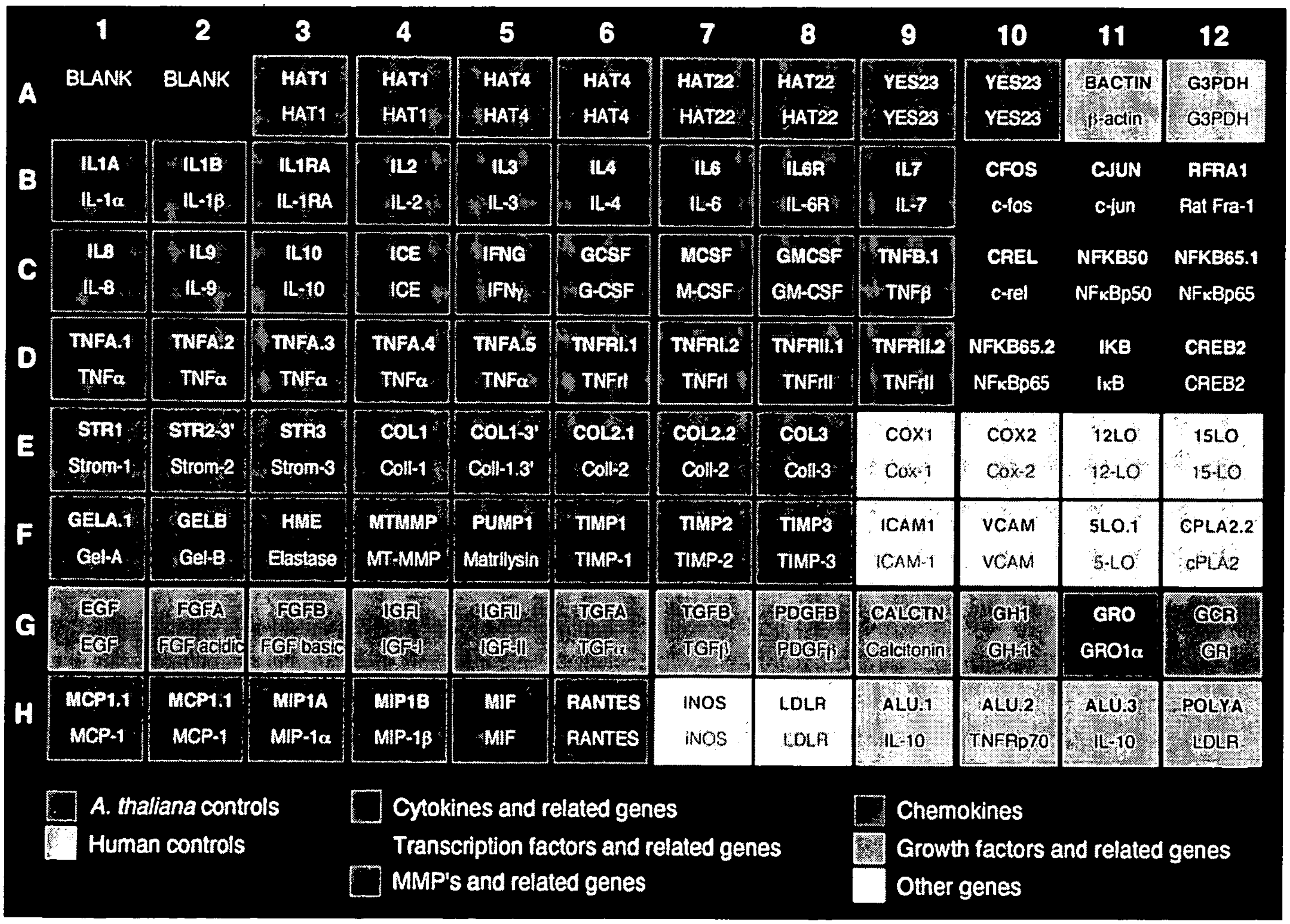
Cultured Cells. The Mono Mac-6 (MM6) monocytic cells (8) were grown in RPMI medium. Human chondrosarcoma SW1353 cells, primary human chondrocytes, and synoviocytes (9, 10) were cultured in DMEM; all culture media were supplemented with 10% fetal bovine serum, 100 μg/ml streptomycin, and 500 units/ml penicillin. Treatment of cells with lipopolysaccharide (LPS) endotoxin at 30 ng/ml, phorbol 12-myristate 13-acetate (PMA) at 50 ng/ml, tumor necrosis factor α (TNF-α) at 50 ng/ml, interleukin (IL)-1β at 30 ng/ml, or transforming growth factor-β (TGF-β) at 100 ng/ml is described in the figure legends.

Fluorescent Probe, Hybridization, and Scanning. Isolation of mRNA, probe preparation, and quantitation with *Arabidopsis* control mRNAs was essentially as described (3) except for the following minor modification. Following the reverse transcriptase step, the appropriate Cy3- and Cy5-labeled samples were pooled; mRNA degraded by heating the sample to 65°C for 10 min with the addition of 5 μl of 0.5M NaOH plus 0.5 ml of 10 mM EDTA. The pooled cDNA was purified from unincorporated nucleotides by gel filtration in Centri-spin columns (Princeton Separations, Adelphia, NJ). Samples were lyophilized and dissolved in 6 μl of hybridization buffer (5× SSC plus 0.2% SDS). Hybridizations, washes, scanning, quantitation procedures, and pseudocolor representations of fluorescent images have been described (3). Scans for the two fluorescent probes were normalized either to the fluorescence intensity of *Arabidopsis* mRNAs spiked into the labeling reactions (see Figs. 2–4) or to the signal intensity of β-actin and glyceraldehyde-3-phosphate dehydrogenase (GAPDH; see Fig. 5).

RESULTS

Ninety-Six-Gene Microarray Design. The actions of cytokines, growth factors, chemokines, transcription factors, MMPs, prostaglandins, and leukotrienes are well recognized in inflammatory disease, particularly RA (11–14). Fig. 1 displays the selected genes for this study and also includes control cDNAs of housekeeping genes such as β-actin and GAPDH and genes from *Arabidopsis* for signal normalization and quantitation (row A, columns 1–12).

Defining Microarray Assay Conditions. Different lengths and concentrations of target DNA were tested by arraying PCR-



	1	2	3	4	5	6	7	8	9	10	11	12
A	BLANK	BLANK	HAT1 HAT1	HAT1 HAT1	HAT4 HAT4	HAT4 HAT4	HAT22 HAT22	HAT22 HAT22	YES23 YES23	YES23 YES23	BACTIN β-actin	G3PDH G3PDH
B	IL1A IL-1α	IL1B IL-1β	IL1RA IL-1RA	IL2 IL-2	IL3 IL-3	IL4 IL-4	IL6 IL-6	IL6R IL-6R	IL7 IL-7	CFOS c-fos	CJUN c-jun	RFRA1 Rat Fra-1
C	IL8 IL-8	IL9 IL-9	IL10 IL-10	ICE ICE	IFNG IFNγ	GCSF G-CSF	MCSF M-CSF	GMCSF GM-CSF	TNFB.1 TNFβ	CREL c-rel	NFKB50 NFκBp50	NFKB65.1 NFκBp65
D	TNFA.1 TNFα	TNFA.2 TNFα	TNFA.3 TNFα	TNFA.4 TNFα	TNFA.5 TNFα	TNFRI.1 TNFrI	TNFRI.2 TNFrI	TNFRII.1 TNFrII	TNFRII.2 TNFrII	NFKB65.2 NFκBp65	IKB IκB	CREB2 CREB2
E	STR1 Strom-1	STR2-3' Strom-2	STR3 Strom-3	COL1 Coll-1	COL1-3' Coll-1.3'	COL2.1 Coll-2	COL2.2 Coll-2	COL3 Coll-3	COX1 Cox-1	COX2 Cox-2	12LO 12-LO	15LO 15-LO
F	GELA.1 Gel-A	GELB Gel-B	HME Elastase	MTMMP MT-MMP	PUMP1 Matrilysin	TIMP1 TIMP-1	TIMP2 TIMP-2	TIMP3 TIMP-3	ICAM1 ICAM-1	VCAM VCAM	5LO.1 5-LO	CPLA2.2 cPLA2
G	EGF EGF	FGFA FGF acidic	FGFB FGF basic	IGFI IGF-I	IGFII IGF-II	TGFA TGFα	TGFB TGFβ	PDGFB PDGFβ	CALCTN Calcitonin	GH1 GH-1	GRO GRO1α	GCR GR
H	MCP1.1 MCP-1	MCP1.1 MCP-1	MIP1A MIP-1α	MIP1B MIP-1β	MIF MIF	RANTES RANTES	INOS iNOS	LDLR LDLR	ALU.1 IL-10	ALU.2 TNFRp70	ALU.3 IL-10	POLYA LDLR

▨ *A. thaliana* controls
□ Human controls
■ Cytokines and related genes
Transcription factors and related genes
■ MMP's and related genes
■ Chemokines
▨ Growth factors and related genes
□ Other genes

FIG. 1. Ninety-six-element microarray design. The target element name and the corresponding gene are shown in the layout. Some genes have more than one target element to guarantee specificity of signal. For TNF the targets represent decreasing lengths of 1, 0.8, 0.6, 0.4, and 0.2 kb from left to right.

amplified products ranging from 0.2 to 1.2 kb at concentrations of 1 μg/μl or less. No significant difference in the signal levels was observed within this range of target size and only with 0.2-kb length was a signal reduced upon an 8-fold dilution of the 1 μg/μl sample (data not shown). In this study the average length of the targets was 1 kb, with a few exceptions in the range of ≈300 bp, arrayed at a concentration of 1 μg/μl. Normally one PCR provided sufficient material to fabricate up to 1000 microarray targets.

In considering positional effects in the development of the targets for the microarrays, selection was biased toward the 3′ proximal regions, because the signal was reduced if the target fragment was biased toward the 5′ end (data not shown). This result was anticipated since the hybridizing probe is prepared by reverse transcription with oligo(dT)-primed mRNA and is richer in 3′ proximal sequences. Cross-hybridizations of probes to targets of a gene family were analyzed with the matrix metalloproteinases as the example because they can show regions of sequence identities of greater than 70%. With collagenase-1 (Col-1) and collagenase-2 (Col-2) genes as targets with up to 70% sequence identity, and stromelysin-1 (Strom-1) and stromelysin-2 (Strom-2) genes with different degrees of identity, our results showed that a short region of overlap, even with 70–90% sequence identity, produced a low level of cross-hybridization. However, shorter regions of identity spread over the length of the target resulted in cross-hybridization (data not shown). For closely related genes, targets were designed by avoiding long stretches of homology. For members of a gene family two or more target regions were included to discriminate between specificity of signal versus cross-hybridization.

Monitoring Differential Expression in Cultured Cell Lines. In RA tissue, the monocyte/macrophage population plays a prominent role in phagocytic and immunomodulatory activities. Typ-

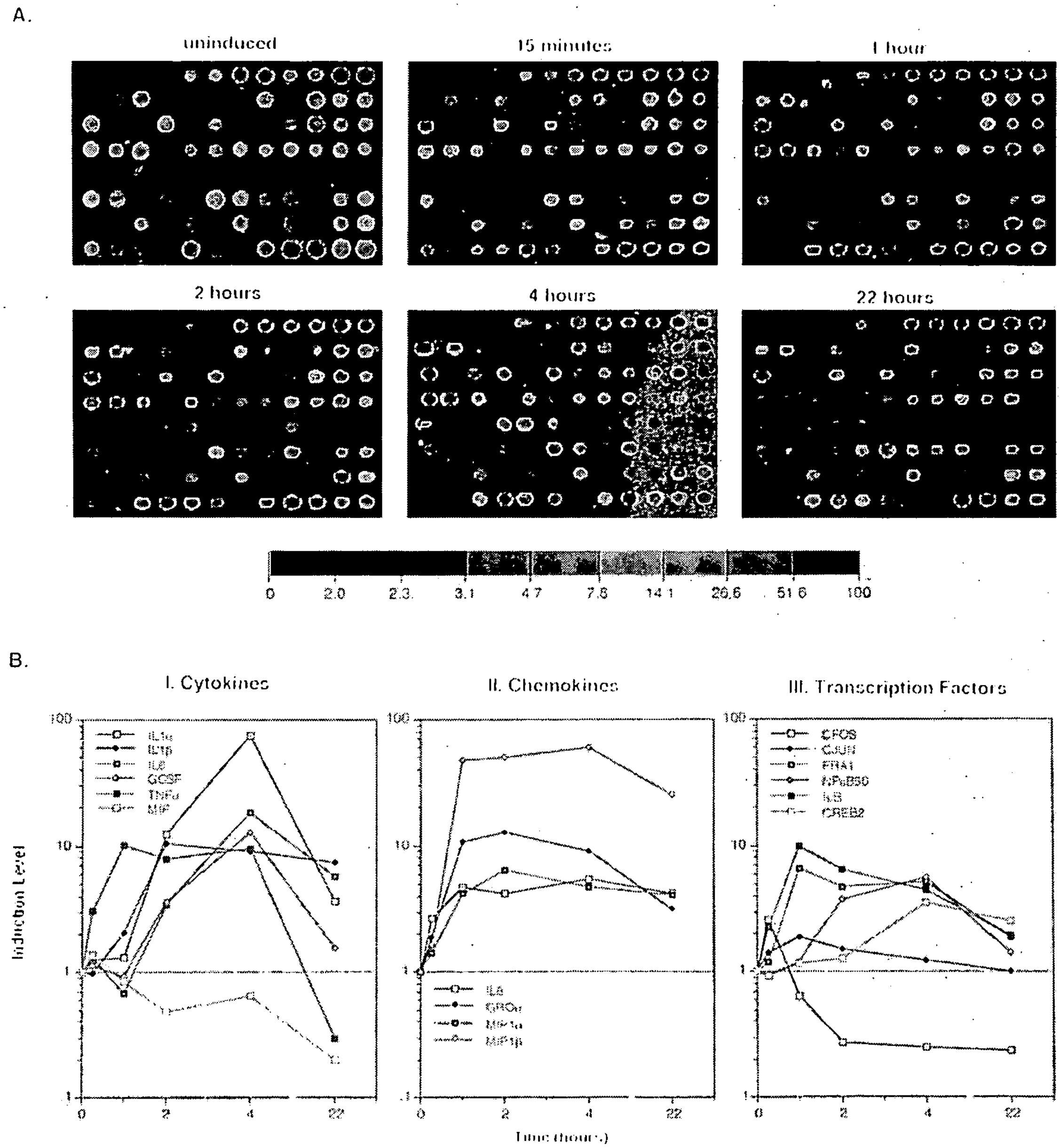


FIG. 2. Time course for LPS/PMA-induced MM6 cells. Array elements are described in Fig. 1. (*A*) Pseudocolor representations of fluorescent scans correspond to gene expression levels at each time point. The array is made up of 8 *Arabidopsis* control targets and 86 human cDNA targets, the majority of which are genes with known or suspected involvement in inflammation. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation. Fluorescent probes were made by labeling mRNA from untreated MM6 cells or LPS and PMA treated cells. mRNA was isolated at indicated times after induction. (*B I–III*) The two-color samples were cohybridized, and microarray scans provided the data for the levels of select transcripts at different time points relative to abundance at time zero. The analysis was performed using normalized data collected from 8-bit images.

ically these cells, when triggered by an immunogen, produce the proinflammatroy cytokines TNF and IL-1. We have used the monocyte cell line MM6 and monitored changes in gene expression upon activation with LPS endotoxin, a component of Gram-negative bacterial membranes, and PMA, which augments the action of LPS on TNF production (15). RNA was isolated at different times after induction and used for cDNA probe preparation. From this time course it was clear that TNF expression was induced within 15 min of treatment, reached maximum levels in 1 hr, remained high until 4 hr and subsequently declined (Fig. 2*A*). Many other cytokine genes were also transiently activated, such as IL-1α and -β, IL-6, and granulocyte colony-stimulating factor (GCSF). Prominent chemokines activated were IL-8, macrophage inflammatory protein (MIP)-1β, more so than MIP-1α, and Groα or melanoma growth stimulatory factor. Migration inhibitory factor (MIF) expressed in the uninduced state declined in LPS-activated cells. Of the immediate early genes, the noticeable ones were *c-fos*, *fra-1*, *c-jun*, NF-κBp50, and IκB, with *c-rel* expression observed even in the uninduced state (Fig. 2*B*). These expression patterns are consistent with reported patterns of activation of certain LPS- and PMA-induced genes (12). Demonstrated here is the unique ability of this system to allow parallel visualization of a large number of gene activities over a period of time.

SW1353 cells is a line derived from malignant tumors of the cartilage and behaves much like the chondrocytes upon stimulation with TNF and IL-1 in the expression of MMPs (9). In addition to confirming our earlier observations with Northern blots on Strom-1, Col-1, and Col-3 expression (9), gelatinase (Gel) A, putative metalloproteinase (PUMP)-1 membrane-type matrix metalloproteinase, tissue inhibitors of matrix metalloproteinases or tissue inhibitor of metalloproteinase 1 (TIMP-1), -2, and -3 were also expressed by these cells together with the human matrix metallo-elastase (HME; Fig. 3*A*). HME induction was estimated to be ≈50-fold and was greater than any of the other MMPs examined (Fig. 3*B*). This result was unexpected because HME is reportedly expressed only by alveolar macrophage and placental cells (16). Expression of the cytokines and chemokines, IL-6, IL-8, MIF, and MIP-1β was also noted. A variety of other genes, including certain transcription factors, were also up-regulated (Fig. 3), but the overall time-dependent expression of genes in the SW1353 cells was qualitatively distinct from the MM6 cells.

Quantitation of differential gene expression (Figs. 2*B* and 3*B*) was achieved with the simultaneous hybridization of Cy3-labeled cDNA from untreated cells and Cy5-labeled cDNA from treated samples. The estimated increases in expression from these microarrays for a select number of genes including IL-1β, IL-8, MIP-1β, TNF, HME, Col-1, Col-3, Strom-1, and Strom-2 were compared with data collected from dot blot analysis. Results (not shown) were in close agreement and confirmed our earlier observations on the use of the microarray method for the quantitation of gene expression (3).

Expression Profiles in Primary Chondrocytes and Synoviocytes of Human RA Tissue. Given the sensitivity and the specificity of this method, expression profiles of primary synoviocytes and chondrocytes from diseased tissue were examined. Without prior exposure to inducing agents, low level expression of *c-jun*, GCSF, IL-3, TNF-β, MIF, and RANTES (regulated upon activation, normal T cell expressed and secreted) was seen as well as expression of MMPs, GelA, Strom-1, Col-1, and the three TIMPs. In this case, Col-2 hybridization was considered to be nonspecific because the second Col-2 target taken from the 3′ end of the gene gave no

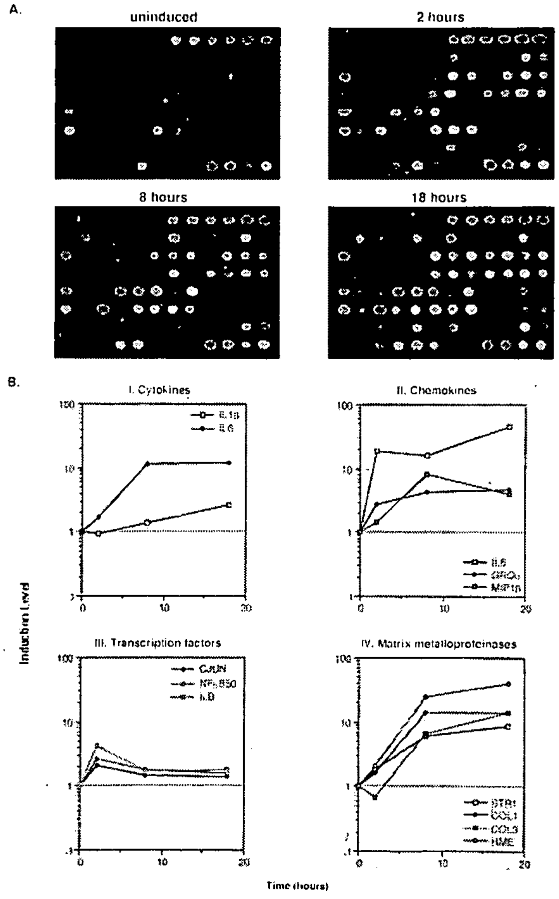


FIG. 3. Time course for IL-1β and TNF-induced SW1353 cells using the inflammation array (Fig. 1). (*A*) Pseudocolor representation of fluorescent scans correspond to gene expression levels at each time point. (*B I–IV*) Relative levels of selected genes at different time points compared with time zero.

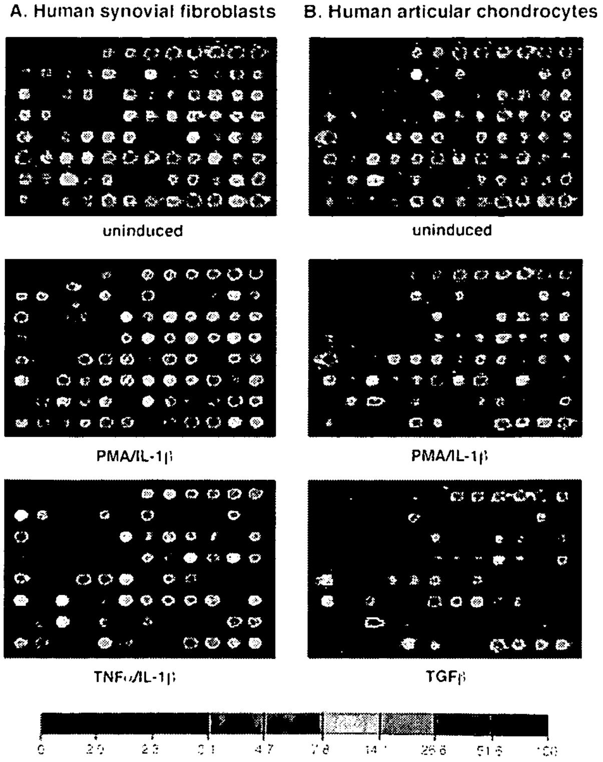


FIG. 4. Expression profiles for early passage primary synoviocytes and chondrocytes isolated from RA tissue, cultured in the presence of 10% fetal calf serum and activated with PMA and IL-1β, or TNF and IL-1β, or TGF-β for 18 hr. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation

signal. Treatment more so with PMA and IL-1, than TNF and IL-1, produced a dramatic up-regulation in expression of several genes in both of these primary cell types. These genes are as follows: the cytokine IL-6, the chemokines IL-8 and Gro-1α, and the MMPs; Strom-1, Col-1, Col-3, and HME; and the adhesion molecule, vascular cell adhesion molecule 1 (VCAM-1). The surprise again is HME expression in these primary cells, for reasons discussed above. From these results, the expression profiles of synoviocytes and the chondrocytes appear very similar; the differences are more quantitative than qualitative. Treatment of the primary chondrocytes with the anabolic growth factor TGF-β had an interesting profile in that it produced a remarkable down-regulation of genes expressed in both the untreated and induced state (Fig. 4).

Given the demonstrated effectiveness of this technology, a comparative analysis of two different inflammatory disease states was conducted with probes made from RA tissue and IBD samples. RA samples were from late stage rheumatoid synovial tissue, and IBD specimens were obtained from inflamed lower intestinal mucosa of patients with Crohn disease. With both the 96-element known gene microarray and the 1000-gene microarray of cDNAs selected from a peripheral human blood cell library (3), distinct differences in gene expression patterns were evident. On the 96-gene array, RA tissue samples from different affected individuals gave similar profiles (data not shown) as did different samples from the same individual (Fig. 5). These patterns were notably similar to those observed with primary synoviocytes and chondrocytes (Fig. 4). Included in the list of prominently up-regulated genes are IL-6, the MMPs Strom-1, Col-1, GelA, HME, and in certain samples PUMP, TIMPs, particularly TIMP-1 and TIMP-3, and the adhesion molecule VCAM. Discernible levels of macrophage chemotactic protein 1 (MCP-1), MIF and RANTES were also noted. IBD samples were in comparison, rather subdued although IL-1 converting enzyme (ICE), TIMP-1, and MIF were notable in all the three different IBD samples examined here. In IBD-A, one of three individual samples, ICE, VCAM, Groα, and MMP expression was more pronounced than in the others.

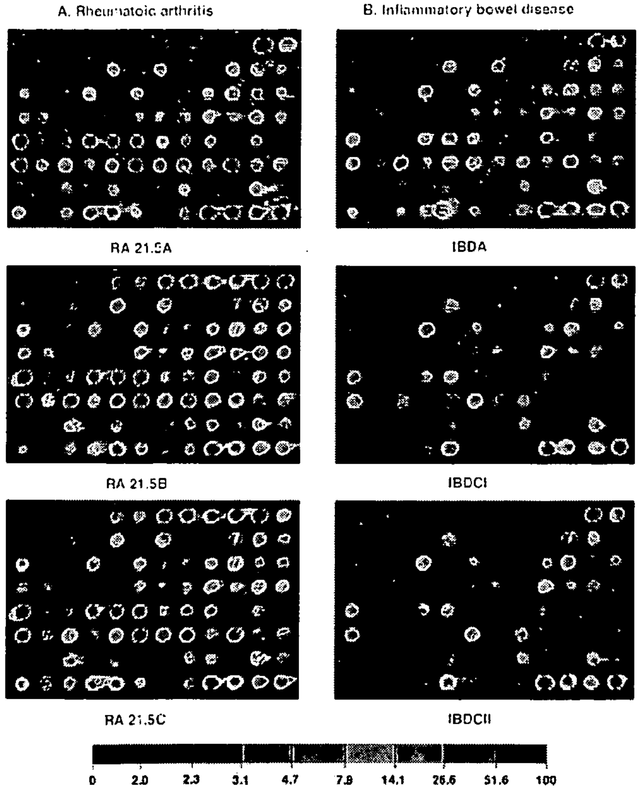


FIG. 5. Expression profiles of RA tissue (*A*) and IBD tissue (*B*). mRNA from RA tissue samples obtained from the same individual was isolated directly after excision (RA 21.5A) or maintained in culture without serum for 2 hr (RA 21.5B) or for 6 hr (RA 21.5C). Profiles from tissue samples of two other individuals (data not shown) were remarkably similar to the ones shown here. IBD-A and IBD-CI are from mRNA samples prepared directly after surgery from two separate individuals. For the IBD-CII probe, the tissue sample was cultured in medium without serum for 2 hr before mRNA preparation.

We also made use of a peripheral blood cDNA library (3) to identify genes expressed by lymphocytes infiltrating the inflamed tissues from the circulating blood. With the 1046-element array of randomly selected cDNAs from this library, probes made from RA and IBD samples showed hybridizations to a large number of genes. Of these, many were common between the two disease tissues while others were differentially expressed (data not shown). A complete survey of these genes was beyond the scope of this study, but for this report we picked three genes that were up-regulated in the RA tissue relative to IBD. These cDNAs were sequenced and identified by comparison to the GenBank database. They are TIMP-1, apoferritin light chain, and manganese superoxide dismutase (MnSOD). Differential expression of MnSOD was only observed in samples of RA tissue explants maintained in growth medium without serum for anywhere between 2 to 16 hr. These results also indicate that the expression profile of genes can be altered when explants are transferred to culture conditions.

DISCUSSION

The speed, ease, and feasibility of simultaneously monitoring differential expression of hundreds of genes with the cDNA microarray based system (1–3) is demonstrated here in the analysis of a complex disease such as RA. Many different cell types in the RA tissue; macrophages, lymphocytes, plasma cells, neutrophils, synoviocytes, chondrocytes, etc. are known to contribute to the development of the disease with the expression of gene products known to be proinflammatory. They include the cytokines, chemokines, growth factors, MMPs, eicosanoids, and others (7, 11–14), and the design of the 96-element known gene microarray was based on this knowledge and depended on the availability of the genes. The technology was validated by confirming earlier observations on the expression of TNF by the monocyte cell line MM6, and of Col-1 and Col-3 expression in the chondrosarcoma cells and articular chondrocytes (9, 12). In our time-dependent survey the chronological order of gene activities in and between gene families was compared and the results have provided unprecedented profiles of the cytokines (TNF, IL-1, IL-6, GCSF, and MIF), chemokines (MIP-1α, MIP-1β, IL-8, and Gro-1), certain transcription factors, and the matrix metalloproteinases (GelA, Strom-1, Col-1, Col-3, HME) in the macrophage cell line MM6 and in the SW1353 chondrosarcoma cells.

Earlier reports of cytokine production in the diseased state had established a model in which TNF is a major participant in RA. Its expression reportedly preceded that of the other cytokines and effector molecules (4). Our results strongly support these results as demonstrated in the time course of the MM6 cells where TNF induction preceded that of IL-1α and IL-β followed by IL-6 and GCSF. These expression profiles demonstrate the utility of the microarrays in determining the hierarachy of signaling events.

In the SW1353 chondrosarcoma cells, all the known MMPs and TIMPs were examined simultaneously. HME expression was discovered, which previously had been observed in only the stromal cells and alveolar macrophages of smoker's lungs and in placental tissue. Its presence in cells of the RA tissue is meaningful because its activity can cause significant destruction of elastin and basement membrane components (16, 17). Expression profiles of synovial fibroblasts and articular chondrocytes were remarkably similar and not too different from the SW1353 cells, indicating that the fibroblast and the chondrocyte can play equally aggressive roles in joint erosion. Prominent genes expressed were the MMPs, but chemokines and cytokines were also produced by these cells. The effect of the anabolic growth factor TGF-β was profoundly evident in demonstrating the down regulation of these catabolic activities.

RA tissue samples undeniably reflected profiles similar to the cell types examined. Active genes observed were IL-3, IL-6, ICE, the MMPs including HME and TIMPs, chemokines IL-8, Groα, MIP, MIF, and RANTES, and the adhesion molecule VCAM. Of the growth factors, fibroblast growth factor β was observed most frequently. In comparison, the expression patterns in the other inflammatory state (i.e., IBD) were not as marked as in the RA samples, at least as obtained from the tissue samples selected for this study.

As an alternative approach, the 1046 cDNA microarray of randomly selected genes from a lymphocyte library was used to identify genes expressed in RA tissue (3). Many genes on this array hybridized with probes made from both RA and IBD tissue samples. The results are not surprising because inflammatory tissue is abundantly supplied with cell types infiltrating from the circulating blood, made apparent also by the high levels of chemokine expression in RA tissue. Because of the magnitude of the effort required to identify all the hybridized genes, we have for this report chosen to describe only three differentially expressed genes mainly to verify this method of analysis.

Of the large number of genes observed here, a fair number were already known as active participants in inflammatory disease. These are TNF, IL-1, IL-6, IL-8, GCSF, RANTES, and VCAM. The novel participants not previously reported are HME, IL-3, ICE, and Groα. With our discovery of HME expression in RA, this gene becomes a target for drug intervention. ICE is a cysteine protease well known for its IL-1β processing activity (18), and recognized for its role in apoptotic cell death (19). Its expression in RA tissue is intriguing. IL-3 is recognized for its growth-promoting activity in hematopoietic cell lineages, is a product of activated T cells (20), and its expression in synoviocytes and chondrocytes of RA tissue is a novel observation.

Like IL-8, Groα, is a C-X-C subgroup chemokine and is a potent neutrophil and basophil chemoattractant. It downregulates the expression of types I and III interstitial collagens (21, 22) and is seen here produced by the MM6 cells, in primary synoviocytes, and in RA tissue. With the presence of RANTES, MCP, and MIP-1β, the C-C chemokines (23) migration and infiltration of monocytes, particularly T cells, into the tissue is also enhanced (5) and aid in the trafficking and recruitment of leukocytes into the RA tissue. Their activation, phagocytosis, degranulation, and respiratory bursts could be responsible for the induction of MnSOD in RA. MnSOD is also induced by TNF and IL-1 and serves a protective function against oxidative damage. The induction of the ferritin light chain encoding gene in this tissue may be for reasons similar to those for MnSOD. Ferritin is the major intracellular iron storage protein and it is responsive to intracellular oxidative stress and reactive oxygen intermediates generated during inflammation (24, 25). The active expression of TIMP-1 in RA tissue, as detected by the 1000-element array, is no surprise because our results have repeatedly shown TIMP-1 to be expressed in the constitutive and induced states of RA cells and tissues.

The suitability of the cDNA microarray technology for profiling diseases and for identifying disease related genes is well documented here. This technology could provide new targets for drug development and disease therapies, and in doing so allow for improved treatment of chronic diseases that are challenging because of their complexity.

We would like to thank the following individuals for their help in obtaining reagents or providing cDNA clones to use as templates in target preparation: N. Arai, P. Cannon, D. R. Cohen, T. Curran, V. Dixit, D. A. Geller, G. I. Goldberg, M. Karin, M. Lotz, L. Matrisian, G. Nolan, C. Lopez-Otin, T. Schall, S. Shapiro, I. Verma, and H. Van Wart. Support for R.W.D., M.S., and R.A.H. was provided by the National Institutes of Health (Grants R37HG00198 and HG00205).

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270**, 467–470.
2. Shalon, D., Smith, S. & Brown, P. O. (1996) *Genome Res.* **6**, 639–645.
3. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 10614–10619.
4. Feldmann, M., Brennan F. M. & Maini, R. N. (1996) *Rheumatoid Arthritis Cell* **85**, 307–310.
5. Schall, T. J. (1994) in *The Cytokine Handbook*, ed. Thomson, A. W. (Academic, New York), 2nd Ed., pp. 410–460.
6. Lotz, M. F., Blanco, J., Von Kempis, J., Dudler, J., Maier, R., Villiger P. M. & Geng, Y. (1995) *J. Rheumatol.* **22**, Supplement 43, 104–108.
7. Birkedal-Hansen, H., Moore, W. G. I., Bodden, M. K., Windsor, L. J., Birkedal-Hansen, B., DeCarlo, A. &. Engler, J. A. (1993) *Crit. Rev. Oral Biol. Med.* **4**, 197–250.
8. Zeigler-Heitbrock, H. W. L., Thiel, E., Futterer, A., Volker, H., Wirtz, A. & Reithmuller, G. (1988) *Int. J. Cancer* **41**, 456–461.
9. Borden, P., Solymar, D., Sucharczuk, A., Lindman, B., Cannon, P. & Heller, R. A. (1996) *J. Biol. Chem.* **271**, 23577–23581.
10. Gadher, S. J. & Woolley, D. E. (1987) *Rheumatol. Int.* **7**, 13–22.
11. Harris, E. D., Jr. (1990) *New Engl. J. Med.* **322**, 1277–1289.
12. Firestein, G. S. (1996) in *Textbook of Rheumatology*, eds. Kelly, W. N., Harris, E. D., Ruddy, S. & Sledge, C. B. (Saunders, Philadelphia), 5th Ed. pp. 5001–5047.
13. Alvaro-Garcia, J. M., Zvaifler, Nathan J., Brown, C. B., Kaushansky, K. & Firestein, Gary S. (1991) *J. Immunol.* **146**, 3365–3371.
14. Firestein, G. S., Alvaro-Grarcia, J. M. & Maki, R. (1990) *J. Immunol.* **144**, 3347–3352.
15. Pradines-Figueres, A. & Raetz, C. R. H. (1992) *J. Biol. Chem.* **267**, 23261–23268.
16. Shapiro, S. D., Kobayashi, D. L. & Ley, T. J. (1993) *J. Biol. Chem.* **208**, 23824–23829.
17. Shipley, M. J., Wesselschmidt, R. L., Kobayashi, D. K., Ley, T. J. & Shapiro, S. D. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 3042–3946.
18. Cerreti, D. P., Kozlosky, C. J., Mosley, B., Nelson, N., Van Ness, K., Greenstreet, T. A., March, C. J., Kronheim, S. R., Druck, T., Cannizaro, L. A., Huebner, K. & Black, R. A. (1992) *Science* **256**, 97–100.
19. Miura, M., Zhu, H., Rotello, R., Hartweig, E. A. & Yuan, J. (1993) *Cell* **75**, 653–660.
20. Arai, K., Lee, F., Miyajima, A., Shoichiro, M., Arai, N. & Takashi, Y. (1990) *Annu. Rev. Biochem.* **59**, 783–836.
21. Geiser, T., Dewald, B., Ehrengruber, M. U., Lewis, I. C. & Baggiolini, M. (1993) *J. Biol. Chem.* **268**, 15419–15424.
22. Unemori, E. N., Amento, E. P., Bauer, E. A. & Horuk, R. (1993) *J. Biol. Chem.* **268**, 1338–1342.
23. Robinson, E., Keystone, E. C., Schall, T. J., Gillet, N. & Fish, E. N. (1995) *Clin. Exp. Immunol.* **101**, 398–407.
24. Roeser, H. (1980) in *Iron Metabolism in Biochemistry and Medicine*, eds. Jacobs, A. & Worwood, M. (Academic, New York), Vol. 2, pp. 605–640.
25. Kwak, E. L., Larochelle, D. A., Beaumont, C., Torti, S. V. & Torti, F. M. (1995) *J. Biol. Chem.* **270**, 15285–15293.

Axl1p sequence following Ser^{206} and occurs within the domain of Axl1p that shows homology with hIDE (*14*). To delete the complete *STE23* sequence and create the *ste23Δ::URA3* mutation, polymerase chain reaction (PCR) primers (5'-TCGGAAGACCTCATTCTTGCTCATTTTGATATTGCTC- TGTAGATTGTACTGAGAGTGCAC-3'; and 5'-GCTACAAACAGCGTCGACTTGAATGCCCCGACATCTTCGACTGTGCGGTATTTCACACCG-3') were used to amplify the *URA3* sequence of pRS316, and the reaction product was transformed into yeast for one-step gene replacement [R. Rothstein, *Methods Enzymol.* 194, 281 (1991)]. To create the *axl1Δ::LEU2* mutation contained on p114, a 5.0-kb Sal I fragment from p*AXL1* was cloned into pUC19, and an internal 4.0-kb Hpa I–Xho I fragment was replaced with a *LEU2* fragment. To construct the *ste23Δ::LEU2* allele (a deletion corresponding to 931 amino acids) carried on p153, a *LEU2* fragment was used to replace the 2.8-kb Pml I–Ecl136 II fragment of *STE23*, which occurs within a 6.2-kb Hind III–Bgl II genomic fragment carried on pSP72 (Promega). To create YEp*MFA1*, a 1.6-kb Bam HI fragment containing *MFA1*, from pKK16 [K. Kuchler, R. E. Sterne, J. Thorner, *EMBO J.* 8, 3973 (1989)], was ligated into the Bam HI site of YEp351 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)].

24. J. Chant and I. Herskowitz, *Cell* 65, 1203 (1991).
25. B. W. Matthews, *Acc. Chem. Res.* 21, 333 (1988).
26. K. Kuchler, H. G. Dohlman, J. Thorner, *J. Cell Biol.* 120, 1203 (1993); R. Kolling and C. P. Hollenberg, *EMBO J.* 13, 3261 (1994); C. Berkower, D. Loayza, S. Michaelis, *Mol. Biol. Cell* 5, 1185 (1994).
27. A. Bender and J. R. Pringle, *Proc. Natl. Acad. Sci. U.S.A* 86, 9976 (1989); J. Chant, K. Corrado, J. R. Pringle, I. Herskowitz, *Cell* 65, 1213 (1991); S. Powers, E. Gonzales, T. Christensen, J. Cubert, D. Broek, *ibid.*, p. 1225; H. O. Park, J. Chant, I. Herskowitz, *Nature* 365, 269 (1993); J. Chant, *Trends Genet.* 10, 328 (1994); ______ and J. R. Pringle, *J. Cell Biol.* 129, 751 (1995); J. Chant, M. Mischke, E. Mitchell, I. Herskowitz, J. R. Pringle, *ibid.*, p. 767.
28. G. F. Sprague Jr., *Methods. Enzymol.* 194, 77 (1991).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
30. A W303 1A derivative, SY2625 (*MATa ura3-1 leu2-3, 112 trp1-1 ade2-1 can1-100 sst1Δ mfa2Δ::FUS1-lacZ his3Δ::FUS1-HIS3*), was the parent strain for the mutant search. SY2625 derivatives for the mating assays, secreted pheromone assays, and the pulse-chase experiments included the following strains: Y49 (*ste22-1*), Y115 (*mfa1Δ::LEU2*), Y142 (*axl1::URA3*), Y173 (*axl1Δ::LEU2*), Y220 (*axl1::URA3 ste23Δ::URA3*), Y221 (*ste23Δ::URA3*), Y231 (*axl1Δ::LEU2 ste23Δ::LEU2*), and Y233 (*ste23Δ::LEU2*). *MATα* derivatives of SY2625 included the following strains: Y199 (SY2625 made *MATα*), Y278 (*ste22-1*), Y195 (*mfa1Δ::LEU2*), Y196 (*axl1Δ::LEU2*), and Y197 (*axl1::URA3*). The EG123 (*MATa leu2 ura3 trp1 can1 his4*) genetic background was used to create a set of strains for analysis of bud site selection. EG123 derivatives included the following strains: Y175 (*axl1Δ::LEU2*), Y223 (*axl1::URA3*), Y234 (*ste23Δ::LEU2*), and Y272 (*axl1Δ::LEU2 ste23Δ::LEU2*). *MATα* derivatives of EG123 included the following strains: Y214 (EG123 made *MATα*) and Y293 (*axl1Δ::LEU2*). All strains were generated by means of standard genetic or molecular methods involving the appropriate constructs (*23*). In particular, the *axl1 ste23* double mutant strains were created by crossing of the appropriate *MATa ste23* and *MATα axl1* mutants, followed by sporulation of the resultant diploid and isolation of the double mutant from nonparental di-type tetrads. Gene disruptions were confirmed with either PCR or Southern (DNA) analysis.
31. p129 is a YEp352 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)] plasmid containing a 5.5-kb Sal I fragment of p*AXL1*. p151 was derived from p129 by insertion of a linker at the Bgl II site within *AXL1*, which led to an in-frame insertion of the hemagglutinin (HA) epitope (DQYPYDVPDYA) (*29*) between amino acids 854 and 855 of the *AXL1* product. pC225 is a KS+ (Stratagene) plasmid containing a 0.5-kb Bam HI–Sst I fragment from p*AXL1*. Substitution mutations of the proposed active site of Axl1p were created with the use of pC225 and site-specific mutagenesis involving appropriate synthetic oligonucleotides (*axl1-H68A*, 5'-GTGCTCACAAAGCGCTGCCAAACCGGC-3'; *axl1-E71A*, 5'-AAGAATCATGTGCGCACAAAGGTGCGC-3'; and *axl1-E71D*, 5'-AAGAATCATGTGATCACAAAGGTGCGC-3'). The mutations were confirmed by sequence analysis. After mutagenesis, the 0.4-kb Bam HI–Msc I fragment from the mutagenized pC225 plasmids was transferred into p*AXL1* to create a set of pRS316 plasmids carrying different *AXL1* alleles, p124 (*axl1-H68A*), p130 (*axl1-E71A*), and p132 (*axl1-E71D*). Similarly, a set of HA-tagged alleles carried on YEp352 were created after replacement of the p151 Bam HI–Msc I fragment, to generate p161 (*axl1-E71A*), p162 (*axl1-*
32

N. Davis, T. Favero, C. de Hoog, and S. Kim for comments on the manuscript. Supported by a grant to C.B. from the Natural Sciences and Engineering Research Council of Canada. Support for M.N.A. was from a California Tobacco-Related Disease Research Program postdoctoral fellowship (4FT-0083).

22 June 1995; accepted 21 August 1995

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena,* Dari Shalon,*† Ronald W. Davis, Patrick O. Brown‡

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 *Arabidopsis* genes were made by means of simultaneous, two-color fluorescence hybridization.

The temporal, developmental, topographical, histological, and physiological patterns in which a gene is expressed provide clues to its biological role. The large and expanding database of complementary DNA (cDNA) sequences from many organisms (*1*) presents the opportunity of defining these patterns at the level of the whole genome.

For these studies, we used the small flowering plant *Arabidopsis thaliana* as a model organism. *Arabidopsis* possesses many advantages for gene expression analysis, including the fact that it has the smallest genome of any higher eukaryote examined to date (*2*). Forty-five cloned *Arabidopsis* cDNAs (Table 1), including 14 complete sequences and 31 expressed sequence tags (ESTs), were used as gene-specific targets. We obtained the ESTs by selecting cDNA clones at random from an *Arabidopsis* cDNA library. Sequence analysis revealed that 28 of the 31 ESTs matched sequences in the database (Table 1). Three additional cDNAs from other organisms served as controls in the experiments.

The 48 cDNAs, averaging ~1.0 kb, were amplified with the polymerase chain reaction (PCR) and deposited into individual wells of a 96-well microtiter plate. Each sample was duplicated in two adjacent wells to allow the reproducibility of the arraying and hybridization process to be tested. Samples from the microtiter plate were printed onto glass microscope slides in an area measuring 3.5 mm by 5.5 mm with the use of a high-speed arraying machine (*3*). The arrays were processed by chemical and heat treatment to attach the DNA sequences to the glass surface and denature them (*3*). Three arrays, printed in a single lot, were used for the experiments here. A single microtiter plate of PCR products provides sufficient material to print at least 500 arrays.

Fluorescent probes were prepared from total *Arabidopsis* mRNA (*4*) by a single round of reverse transcription (*5*). The *Arabidopsis* mRNA was supplemented with human acetylcholine receptor (AChR) mRNA at a dilution of 1:10,000 (w/w) before cDNA synthesis, to provide an internal standard for calibration (*5*). The resulting fluorescently labeled cDNA mixture was hybridized to an array at high stringency (*6*) and scanned

M. Schena and R. W. Davis, Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.
D. Shalon and P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

*These authors contributed equally to this work.
†Present address: Synteni, Palo Alto, CA 94303, USA.
‡To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

with a laser (3). A high-sensitivity scan gave signals that saturated the detector at nearly all of the *Arabidopsis* target sites (Fig. 1A). Calibration relative to the AChR mRNA standard (Fig. 1A) established a sensitivity limit of ~1:50,000. No detectable hybridization was observed to either the rat glucocorticoid receptor (Fig. 1A) or the yeast *TRP4* (Fig. 1A) targets even at the highest scanning sensitivity. A moderate-sensitivity scan of the same array allowed linear detection of the more abundant transcripts (Fig. 1B). Quantitation of both scans revealed a range of expression levels spanning three orders of magnitude for the 45 genes tested (Table 2). RNA blots (7) for several genes (Fig. 2) corroborated the expression levels measured with the microarray to within a factor of 5 (Table 2).

Differential gene expression was investigated with a simultaneous, two-color hybridization scheme, which served to minimize experimental variation inherent in the comparison of independent hybridizations. Fluorescent probes were prepared from two mRNA sources with the use of reverse transcriptase in the presence of fluorescein- and lissamine-labeled nucleotide analogs, respectively (5). The two probes were then mixed together in equal proportions, hybridized to a single array, and scanned separately for fluorescein and lissamine emission after independent excitation of the two fluorophores (3).

To test whether overexpression of a single gene could be detected in a pool of total *Arabidopsis* mRNA, we used a microarray to analyze a transgenic line overexpressing the single transcription factor *HAT4* (8). Fluorescent probes representing mRNA from wild-type and *HAT4*-transgenic plants were labeled with fluorescein and lissamine, respectively; the two probes were then mixed and hybridized to a single array. An intense hybridization signal was observed at the position of the HAT4 cDNA in the lissamine-specific scan (Fig. 1D), but not in the fluorescein-specific scan of the same array (Fig. 1C). Calibration with AChR mRNA added to the fluorescein and lissamine cDNA synthesis reactions at dilutions of 1:10,000 (Fig. 1C) and 1:100 (Fig. 1D), respectively, revealed a 50-fold elevation of *HAT4* mRNA in the transgenic line relative to its abundance in wild-type plants (Table 2). This magnitude of *HAT4* overexpression matched that inferred from the Northern (RNA) analysis within a factor of 2 (Fig. 2 and Table 2). Expression of all the other genes monitored on the array differed by less than a factor of 5 between *HAT4*-transgenic and wild-type plants (Fig 1, C

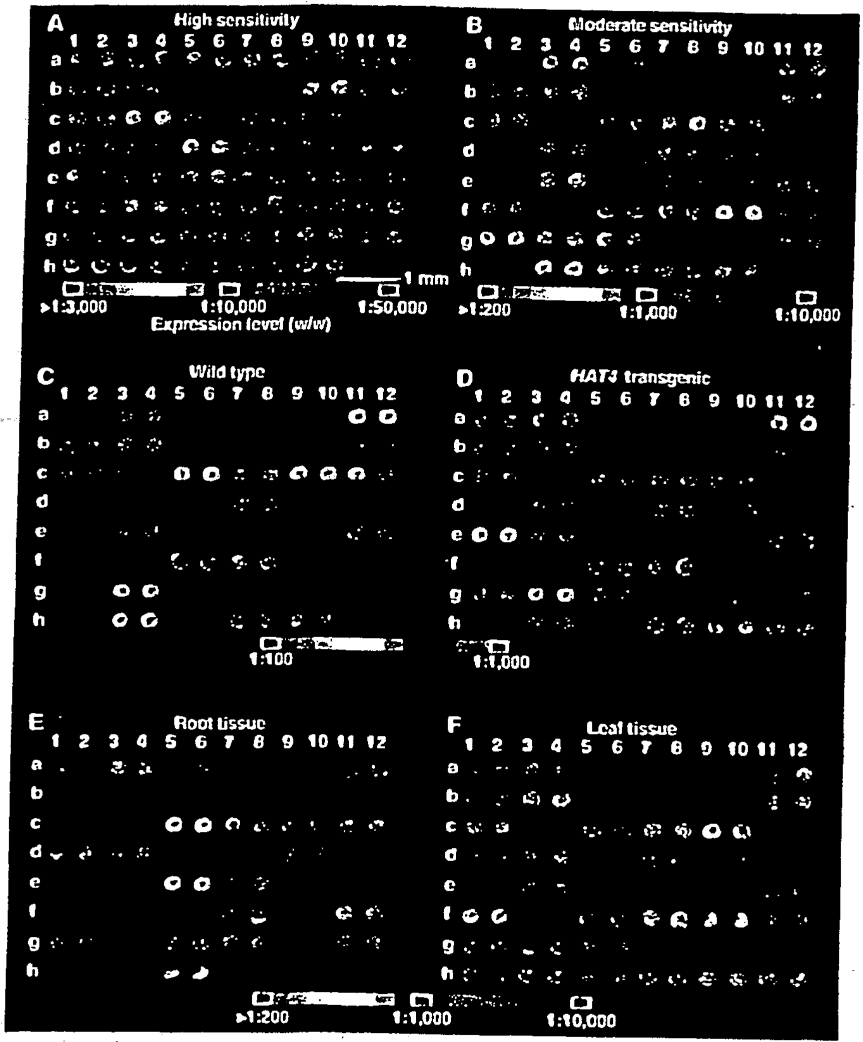


Fig. 1. Gene expression monitored with the use of cDNA microarrays. Fluorescent scans represented in pseudocolor correspond to hybridization intensities. Color bars were calibrated from the signal obtained with the use of known concentrations of human AChR mRNA in independent experiments. Numbers and letters on the axes mark the position of each cDNA. (**A**) High-sensitivity fluorescein scan after hybridization with fluorescein-labeled cDNA derived from wild-type plants. (**B**) Same array as in (A) but scanned at moderate sensitivity. (**C** and **D**) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from wild-type plants and lissamine-labeled cDNA from HAT4-transgenic plants. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNA from wild-type plants (C) and the lissamine fluorescence corresponding to mRNA from HAT4-transgenic plants (D). (**E** and **F**) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from root tissue and lissamine-labeled cDNA from leaf tissue. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNAs expressed in roots (E) and the lissamine fluorescence corresponding to mRNAs expressed in leaves (F).

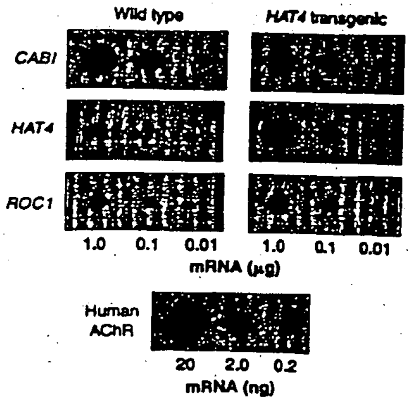


Fig. 2. Gene expression monitored with RNA (Northern) blot analysis. Designated amounts of mRNA from wild-type and *HAT4*-transgenic plants were spotted onto nylon membranes and probed with the cDNAs indicated. Purified human AChR mRNA was used for calibration.

and D, and Table 2). Hybridization of fluorescein-labeled glucocorticoid receptor cDNA (Fig. 1C) and lissamine-labeled *TRP4* cDNA (Fig. 1D) verified the presence of the negative control targets and the lack of optical cross talk between the two fluorophores.

To explore a more complex alteration in expression patterns, we performed a second two-color hybridization experiment with fluorescein- and lissamine-labeled probes prepared from root and leaf mRNA, respectively. The scanning sensitivities for the two fluorophores were normalized by matching the signals resulting from AChR mRNA, which was added to both cDNA synthesis reactions at a dilution of 1:1000 (Fig. 1, E and F). A comparison of the scans revealed widespread differences in gene expression between root and leaf tissue (Fig. 1, E and F). The mRNA from the light-regulated *CABI* gene was ~500-fold more abundant in leaf (Fig. 1F) than in root tissue (Fig. 1E). The expression of 26 other genes differed between root and leaf tissue by more than a factor of 5 (Fig. 1, E and F).

The *HAT4*-transgenic line we examined has elongated hypocotyls, early flowering, poor germination, and altered pigmentation (*8*). Although changes in expression were observed for *HAT4*, large changes in expression were not observed for any of the other 44 genes we examined. This was somewhat surprising, particularly because comparative analysis of leaf and root tissue identified 27 differentially expressed genes. Analysis of an expanded set of genes may be required to identify genes whose expression changes upon *HAT4* overexpression; alternatively, a comparison of mRNA populations from specific tissues of wild-type and *HAT4*-transgenic plants may allow identification of downstream genes.

At the current density of robotic printing, it is feasible to scale up the fabrication process to produce arrays containing 20,000 cDNA targets. At this density, a single array would be sufficient to provide gene-specific targets encompassing nearly the entire repertoire of expressed genes in the *Arabidopsis* genome (*2*). The availability of 20,274 ESTs from *Arabidopsis* (*1*, *9*) would provide a rich source of templates for such studies.

The estimated 100,000 genes in the human genome (*10*) exceeds the number of *Arabidopsis* genes by a factor of 5 (*2*). This modest increase in complexity suggests that similar cDNA microarrays, prepared from the rapidly growing repertoire of human ESTs (*1*), could be used to determine the expression patterns of tens of thousands of human genes in diverse cell types. Coupling an amplification strategy to the reverse transcription reaction (*11*) could make it feasible to monitor expression even in minute tissue samples. A wide variety of acute and chronic physiological and pathological conditions might lead to characteristic changes in the patterns of gene expression in peripheral blood cells or other easily sampled tissues. In concert with cDNA microarrays for monitoring complex expression patterns, these tissues might therefore serve as sensitive in vivo sensors for clinical diagnosis. Microarrays of cDNAs could thus provide a useful link between human gene sequences and clinical medicine.

Table 1. Sequences contained on the cDNA microarray. Shown is the position, the known or putative function, and the accession number of each cDNA in the microarray (Fig. 1). All but three of the ESTs used in this study matched a sequence in the database. NADH, reduced form of nicotinamide adenine dinucleotide; ATPase, adenosine triphosphatase; GTP, guanosine triphosphate.

Position	cDNA	Function	Accession number
a1, 2	AChR	Human AChR	*
a3, 4	EST3	Actin	H36236
a5, 6	EST6	NADH dehydrogenase	Z27010
a7, 8	*AAC1*	Actin 1	M20016
a9, 10	EST12	Unknown	U36594†
a11, 12	EST13	Actin	T45783
b1, 2	*CABI*	Chlorophyll a/b binding	M85150
b3, 4	EST17	Phosphoglycerate kinase	T44490
b5, 6	*GA4*	Gibberellic acid biosynthesis	L37126
b7, 8	EST19	Unknown	U36595†
b9, 10	*GBF-1*	G-box binding factor 1	X63894
b11, 12	EST23	Elongation factor	X52256
c1, 2	EST29	Aldolase	T04477
c3, 4	*GBF-2*	G-box binding factor 2	X63895
c5, 6	EST34	Chloroplast protease	R87034
c7, 8	EST35	Unknown	T14152
c9, 10	EST41	Catalase	T22720
c11, 12	rGR	Rat glucocorticoid receptor	M14053
d1, 2	EST42	Unknown	U36596†
d3, 4	EST45	ATPase	J04185
d5, 6	*HAT1*	Homeobox-leucine zipper 1	U09332
d7, 8	EST46	Light harvesting complex	T04063
d9, 10	EST49	Unknown	T76267
d11, 12	*HAT2*	Homeobox-leucine zipper 2	U09335
e1, 2	*HAT4*	Homeobox-leucine zipper 4	M90394
e3, 4	EST50	Phosphoribulokinase	T04344
e5, 6	*HAT5*	Homeobox-leucine zipper 5	M90416
e7, 8	EST51	Unknown	Z33675
e9, 10	*HAT22*	Homeobox-leucine zipper 22	U09336
e11, 12	EST52	Oxygen evolving	T21749
f1, 2	EST59	Unknown	Z34607
f3, 4	*KNAT1*	*Knotted*-like homeobox 1	U14174
f5, 6	EST60	RuBisCO small subunit	X14564
f7, 8	EST69	Translation elongation factor	T42799
f9, 10	*PPH1*	Protein phosphatase 1	U34803
f11, 12	EST70	Unknown	T44621
g1, 2	EST75	Chloroplast protease	T43698
g3, 4	EST78	Unknown	R65481
g5, 6	*ROC1*	Cyclophilin	L14844
g7, 8	EST82	GTP binding	X59152
g9, 10	EST83	Unknown	Z33795
g11, 12	EST84	Unknown	T45278
h1, 2	EST91	Unknown	T13832
h3, 4	EST96	Unknown	R64816
h5, 6	*SAR1*	Synaptobrevin	M90418
h7, 8	EST100	Light harvesting complex	Z18205
h9, 10	EST103	Light harvesting complex	X03909
h11, 12	*TRP4*	Yeast tryptophan biosynthesis	X04273

*Proprietary sequence of Stratagene (La Jolla, California). †No match in the database; novel EST.

Table 2. Gene expression monitoring by microarray and RNA blot analyses; tg, *HAT4*-transgenic. See Table 1 for additional gene information. Expression levels (w/w) were calibrated with the use of known amounts of human AChR mRNA. Values for the microarray were determined from microarray scans (Fig. 1); values for the RNA blot were determined from RNA blots (Fig. 2).

Gene	Expression level (w/w)	
	Microarray	RNA blot
CABI	1:48	1:83
CABI (tg)	1:120	1:150
HAT4	1:8300	1:6300
HAT4 (tg)	1:150	1:210
ROC1	1:1200	1:1800
ROC1 (tg)	1:260	1:1300

REFERENCES AND NOTES

1. The current EST database (dbEST release 091495) from the National Center for Biotechnology Information (Bethesda, MD) contains a total of 322,225 entries, including 255,645 from the human genome and 21,044 from *Arabidopsis*. Access is available via the World Wide Web (http://www.ncbi.nlm.nih.gov).
2. E. M. Meyerowitz and R. E. Pruitt, *Science* 229, 1214 (1985); R. E. Pruitt and E. M. Meyerowitz, *J. Mol. Biol.* 187, 169 (1986); L Hwang *et al.*, *Plant J.* 1, 367 (1991); P. Jarvis *et al.*, *Plant Mol. Biol.* 24, 685 (1994); L. Le Guen *et al.*, *Mol. Gen. Genet.* 245, 390 (1994).
3. D. Shalon, thesis, Stanford University (1995); ——— and P. O. Brown, in preparation. Microarrays were fabricated on poly-L-lysine–coated microscope slides (Sigma) with a custom-built arraying machine fitted with one printing tip. The tip loaded 1 μl of PCR product (0.5 mg/ml) from 96-well microtiter plates and deposited ~0.005 μl per slide on 40 slides at a spacing of 500 μm. The printed slides were rehydrated for 2 hours in a humid chamber, snap-dried at 100°C for 1 min, rinsed in 0.1% SDS, and treated with 0.05% succinic anhydride prepared in buffer consisting of 50% 1-methyl-2-pyrrolidinone and 50% boric acid. The cDNA on the slides was denatured in distilled water for 2 min at 90°C immediately before use. Microarrays were scanned with a laser fluorescent scanner that contained a computer-controlled XY stage and a microscope objective. A mixed gas, multiline laser allowed sequential excitation of the two fluorophores. Emitted light was split according to wavelength and detected with two photomultiplier tubes. Signals were read into a PC with the use of a 12-bit analog-to-digital board. Additional details of microarray fabrication and use may be obtained by means of e-mail (pbrown@cmgm. stanford.edu).
4. F. M. Ausubel *et al.*, Eds., *Current Protocols in Molecular Biology* (Greene & Wiley Interscience, New York, 1994), pp. 4.3.1–4.3.4.
5. Polyadenylated [poly(A)$^+$] mRNA was prepared from total RNA with the use of Oligotex-dT resin (Qiagen). Reverse transcription (RT) reactions were carried out with a StrataScript RT-PCR kit (Stratagene) modified as follows: 50-μl reactions contained 0.1 μg/μl of *Arabidopsis* mRNA, 0.1 ng/μl of human AChR mRNA, 0.05 μg/μl of oligo(dT) (21-mer), 1× first strand buffer, 0.03 U/μl of ribonuclease block, 500 μM deoxyadenosine triphosphate (dATP), 500 μM deoxyguanosine triphosphate, 500 μM dTTP, 40 μM deoxycytosine triphosphate (dCTP), 40 μM fluorescein-12-dCTP (or lissamine-5-dCTP), and 0.03 U/μl of StrataScript reverse transcriptase. Reactions were incubated for 60 min at 37°C, precipitated with ethanol, and resuspended in 10 μl of TE (10 mM tris-HCl and 1 mM EDTA, pH 8.0). Samples were then heated for 3 min at 94°C and chilled on ice. The RNA was degraded by adding 0.25 μl of 10 N NaOH followed by a 10-min incubation at 37°C. The samples were neutralized by addition of 2.5 μl of 1 M tris-Cl (pH 8.0) and 0.25 μl of 10 N HCl and precipitated with ethanol. Pellets were washed with 70% ethanol, dried to completion in a speedvac, resuspended in 10 μl of H_2O, and reduced to 3.0 μl in a speedvac. Fluorescent nucleotide analogs were obtained from New England Nuclear (DuPont).
6. Hybridization reactions contained 1.0 μl of fluorescent cDNA synthesis product (5) and 1.0 μl of hybridization buffer [10× saline sodium citrate (SSC) and 0.2% SDS]. The 2.0-μl probe mixtures were aliquoted onto the microarray surface and covered with cover slips (12 mm round). Arrays were transferred to a hybridization chamber (3) and incubated for 18 hours at 65°C. Arrays were washed for 5 min at room temperature (25°C) in low-stringency wash buffer (1× SSC and 0.1% SDS), then for 10 min at room temperature in high-stringency wash buffer (0.1× SSC and 0.1% SDS). Arrays were scanned in 0.1× SSC with the use of a fluorescence laser-scanning device (3).
7. Samples of poly(A)$^+$ mRNA (4, 5) were spotted onto nylon membranes (Nytran) and crosslinked with ultraviolet light with the use of a Stratalinker 1800 (Stratagene). Probes were prepared by random priming with the use of a Prime-It II kit (Stratagene) in the presence of [^{32}P]dATP. Hybridizations were carried out according to the instructions of the manufacturer. Quantitation was performed on a PhosphorImager (Molecular Dynamics).
8. M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 89, 3894 (1992); M. Schena, A. M. Lloyd, R. W. Davis, *Genes Dev.* 7, 367 (1993); M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 91, 8393 (1994).
9. H. Hofte *et al.*, *Plant J.* 4, 1051 (1993); T. Newman *et al.*, *Plant Physiol.* 106, 1241 (1994).
10. N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* 88, 7474 (1991); E. D. Green and R. H. Waterston, *J. Am. Med. Assoc.* 266, 1966 (1991); C. Bellanne-Chantelot, *Cell* 70, 1059 (1992); D. R. Cox *et al.*, *Science* 265, 2031 (1994).
11. E. S. Kawasaki *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 85, 5698 (1988).
12. The laser fluorescent scanner was designed and fabricated in collaboration with S. Smith of Stanford University. Scanner and analysis software was developed by R. X. Xia. The succinic anhydride reaction was suggested by J. Mulligan and J. Van Ness of Darwin Molecular Corporation. Thanks to S. Theologis, C. Somerville, K. Yamamoto, and members of the laboratories of R.W.D. and P.O.B. for critical comments. Supported by the Howard Hughes Medical Institute and by grants from NIH [R21HG00450] (P.O.B.) and R37AG00198 (R.W.D.)] and from NSF (MCB9106011) (R.W.D.) and by an NSF graduate fellowship (D.S.). P.O.B. is an assistant investigator of the Howard Hughes Medical Institute.

11 August 1995; accepted 22 September 1995

Gene Therapy in Peripheral Blood Lymphocytes and Bone Marrow for ADA$^-$ Immunodeficient Patients

Claudio Bordignon,* Luigi D. Notarangelo, Nadia Nobili, Giuliana Ferrari, Giulia Casorati, Paola Panina, Evelina Mazzolari, Daniela Maggioni, Claudia Rossi, Paolo Servida, Alberto G. Ugazio, Fulvio Mavilio

Adenosine deaminase (ADA) deficiency results in severe combined immunodeficiency, the first genetic disorder treated by gene therapy. Two different retroviral vectors were used to transfer ex vivo the human ADA minigene into bone marrow cells and peripheral blood lymphocytes from two patients undergoing exogenous enzyme replacement therapy. After 2 years of treatment, long-term survival of T and B lymphocytes, marrow cells, and granulocytes expressing the transferred ADA gene was demonstrated and resulted in normalization of the immune repertoire and restoration of cellular and humoral immunity. After discontinuation of treatment, T lymphocytes, derived from transduced peripheral blood lymphocytes, were progressively replaced by marrow-derived T cells in both patients. These results indicate successful gene transfer into long-lasting progenitor cells, producing a functional multilineage progeny.

Severe combined immunodeficiency associated with inherited deficiency of ADA (*1*) is usually fatal unless affected children are kept in protective isolation or the immune system is reconstituted by bone marrow transplantation from a human leukocyte antigen (HLA)–identical sibling donor (*2*). This is the therapy of choice, although it is available only for a minority of patients. In recent years, other forms of therapy have been developed, including transplants from haploidentical donors (*3*, *4*), exogenous enzyme replacement (*5*), and somatic-cell gene therapy (*6*–*9*).

We previously reported a preclinical model in which ADA gene transfer and expression successfully restored immune functions in human ADA-deficient (ADA$^-$) peripheral blood lymphocytes (PBLs) in immunodeficient mice in vivo (*10*, *11*). On the basis of these preclinical results, the clinical application of gene therapy for the treatment of ADA$^-$ SCID (severe combined immunodeficiency disease) patients who previously failed exogenous enzyme replacement therapy was approved by our Institutional Ethical Committees and by the Italian National Committee for Bioethics (*12*). In addition to evaluating the safety and efficacy of the gene therapy procedure, the aim of the study was to define the relative role of PBLs and hematopoietic stem cells in the long-term reconstitution of immune functions after retroviral vector–mediated ADA gene transfer. For this purpose, two structurally identical vectors expressing the human ADA complementary DNA (cDNA), distinguishable by the presence of alternative restriction sites in a nonfunctional region of the viral long-terminal repeat (LTR), were used to transduce PBLs and bone marrow (BM) cells independently. This procedure allowed identification of the origin of

C. Bordignon, N. Nobili, G. Ferrari, D. Maggioni, C. Rossi, P. Servida, F. Mavilio, Telethon Gene Therapy Program for Genetic Diseases, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
L. D. Notarangelo, E. Mazzolari, A. G. Ugazio, Department of Pediatrics, University of Brescia Medical School, Brescia, Italy.
G. Casorati, Unità di Immunochimica, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
P. Panina, Roche Milano Ricerche, Milan, Italy.

*To whom correspondence should be addressed.